TruckDrive: Long-Range Autonomous Highway Driving Dataset

Filippo Ghilotti[1] Edoardo Palladin[1] Samuel Brucker[1] Adam Sigal[1]
Mario Bijelic[1,2] Felix Heide[1,2]
[1]Torc Robotics [2]Princeton University

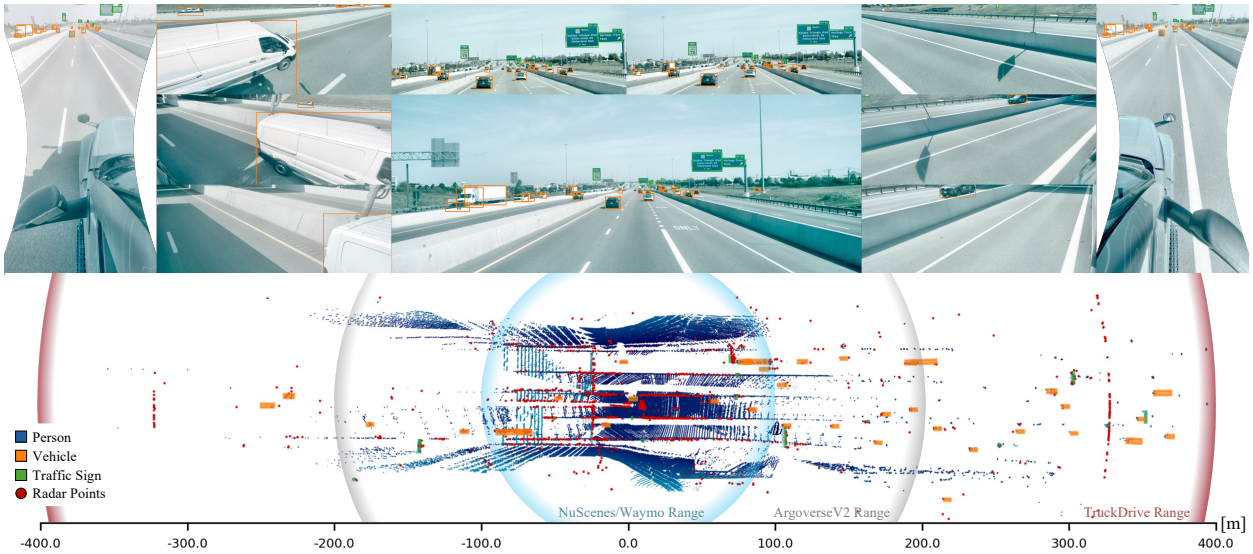


Figure 1. **TruckDrive Dataset.** Autonomous vehicles, especially heavy trucks, require long planning horizons for safe driving in highway scenarios due to higher speed and longer breaking distances. This requires perception ranges well beyond 300 m, while the most common datasets are limited to 100 m [6, 53]. We introduce the TruckDrive Dataset, a large scale multi-modal benchmark captured with a sensor setup tailored for long-range perception with LiDAR, radar and 3D annotations up to 400 m and images and 2D annotations up to 1000 m.

Abstract

Safe highway autonomy for heavy trucks remains an open and unsolved challenge: due to long braking distances, scene understanding of hundreds of meters is required for anticipatory planning and to allow safe braking margins. However, existing driving datasets primarily cover urban scenes, with perception effectively limited to short ranges of only up to 100 meters. To address this gap, we introduce TruckDrive, a highway-scale multimodal driving dataset, captured with a sensor suite purpose-built for long range sensing: seven long-range FMCW LiDARs measuring range and radial velocity, three high-resolution short-range LiDARs, eleven 8MP surround cameras with varying focal lengths and ten 4D FMCW radars. The dataset offers 475 thousands samples with 165 thousands densely annotated frames for driving perception benchmarking up to 1,000 meters for 2D detection and 400 meters for 3D detection, depth estimation, tracking, planning and end to end driving over 20 seconds sequences at highway speeds. We find that state-of-the-art autonomous driving models do not generalize to ranges beyond 150 meters, with drops between 31% and 99% in 3D perception tasks, exposing a systematic long-range gap that current architectures and training signals cannot close. Dataset download, devkit, and videos are available at: light.princeton.edu/TruckDrive.

1. Introduction

Autonomous driving methods require scene understanding, robotic planning and control, either implicitly, in an end-to-end fashion, or in explicit modules, including perception [1, 19, 37, 40, 65], prediction of scene geometry [24, 51] and of the future evolution of relevant agents [23], tracking of the environment [16, 59, 65, 67], planning and control [29, 30].

In the last decade, the development of driving methods have been largely driven by learned models trained on driving datasets including KITTI [21], Cityscapes [14], nuScenes [6], Waymo [53] and Argoverse [10, 60], which predominantly feature urban environments and therefore

All authors contributed equally to this work.

implicitly bias the development of the field toward short-range perception and low-speed driving. This bias is reflected in the annotation range, which typically extends only 70–100 m from the ego-vehicle.

For normal passenger cars driving in urban environments, short-range perception is sufficient as lower speeds convert the limited spatial range into enough temporal foresight to support the 5–10 seconds planning horizons of modern prediction and planning stacks [17, 46, 54, 63]. For heavy-duty trucks operating at highway speeds, instead, the safety envelope is dominated by their high-inertia braking requirements. At 120 km/h, a fully-loaded truck requires over 150–200 m to stop, equivalent to 4.5–6 s of look-ahead perception. Therefore, the necessary braking budget is severely compromised by limited sensing horizons: an 80 m range provides only about 2.4 seconds of foresight, and even 100 m yields merely 3.0 seconds. This entire window is consumed by sensing and planning latencies, eroding the time required for safe braking actuation before the maneuver can even begin. This leaves a critically insufficient margin for the vehicle's deceleration and renders strategic planning, like merging or lane changes, unfeasible.

Driving architectures for long-range perception and planning are non-trivial: Bird's-Eye-View (BEV) and dense voxel representations scale quadratically with distance, leading to exponential growth in compute and memory [47]. Concurrently, the signal-to-noise ratio of distant objects decreases sharply due to sensor resolution limits and atmospheric attenuation. Sparse [61] and range-aware methods [31] alleviate these issues but remain constrained by calibration drift, temporal uncertainty and sparsity of long-range supervision. Moreover, performance on short-range urban benchmarks has begun to saturate, with a decline in the number of submissions and a flattening in the performance gain (Figure 2), with poor generalization capability beyond 100 m [47] of models designed around these priors.

To close this gap, we introduce *TruckDrive*, the first large-scale dataset specifically designed for long-range, high-speed autonomous driving. TruckDrive, as presented in Figure 1, extends the perception range by a factor of *five* relative to urban benchmarks, compared in Table 1, providing 2D annotations up to 1,000 m and corresponding 3D annotations up to 400 m, with 15 to 25 s temporal clips to support forecasting and end-to-end (E2E) learning. The dataset includes over 475 k multi-modal synchronized samples, among which 165 k manually labeled and 310 k unlabeled for self-supervised and unsupervised research. Our sensor suite integrates high-resolution (8MP) short and long focal length cameras, wide-baseline stereo, short and long range 4D LiDARs and 4D radars, enabling comprehensive research in perception, prediction and planning.

We evaluate state-of-the-art driving methods for urban datasets in diverse tasks and observe drops between 31 and

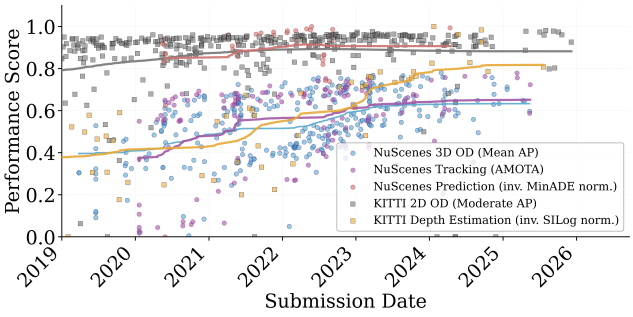


Figure 2. **Performance Saturation on Urban Datasets.** We plot the performance of 2D and 3D OD, Tracking, Prediction and Depth Estimation of NuScenes [6] and Kitti [3] leader boards across the years and observe a saturation of these benchmarks.

99% in 3D perception tasks beyond 150 m, confirming that they do not generalize to long-range regimes. This exposes a fundamental open challenge in current architectures and motivates new directions in efficient representation learning, sensor fusion and long-horizon reasoning.

We make the following contributions:

- We present a long-range, high-fidelity multi-modal driving dataset that combines high-resolution 8MP cameras, large-baseline stereo, 4D LiDARs and 4D radars, enabling dense 3D annotations up to 400 m and 2D annotations up to 1 km.
- We provide large-scale data comprising 475 k samples, including 165 k labeled and 310 k unlabeled frames, with full raw sensor streams to support supervised, semi-supervised, and self-supervised research.
- We establish a highway-scale benchmark for perception, prediction, planning and E2E driving tasks under high-speed, long-range conditions, finding several failure modes and scaling challenges in existing models.

2. Related Work

Public vision datasets [9, 15, 18, 39, 57, 68, 69] have been a catalyst for progress in computer vision, providing a shared basis for developing and comparing novel algorithms. Autonomous driving has followed the same pattern where improvements in detection, prediction and planning have been tightly coupled to increasingly capable datasets.

Early Autonomous Driving Datasets. The field was pioneered by the KITTI dataset [21] and, later, its extensions [2, 3, 38, 43, 44], among the firsts to provide synchronized camera and LiDAR data with $3D$ bounding boxes annotations. These datasets, however, are limited to a relatively small scale, ranges and scenarios.

Large-Scale Multimodal Autonomous Driving Datasets. The next generation of datasets addressed these limitations by introducing 360 degrees sensor coverage and a much larger scale. The nuScenes ecosystem [6] provides a full sensor suite for 3D perception, which was later comple-

Table 1. **TruckDrive Benchmark Comparison.** Cross-dataset summary of sensors, synced samples and useful ranges. TruckDrive couples 7 long range and 3 short range LiDARs with 10 automotive radars, 9 wide/medium field of view cameras and $1-3$ long-focal-length wide-baseline stereo cameras. It offers 165 thousands annotated samples and additional 310 thousands unlabeled samples and extends the effective perception range to $[-400,+400]$ meters, focusing on highway long-range scenarios to stress perception capabilities beyond conventional benchmarks. *NuPlan [7] provides auto-labeled annotations.

Dataset	LiDARs	Radars	Cameras	Localization	Sensor Count	Synced Samples	Manually Annotated	Effective Range
KITTI [21]	1x 64-beam	-	2x RGB, 2x grayscale	GPS, IMU	7	216k	15k	[0, +70]
SemanticKITTI [2]	1x 64-beam	-	2x RGB, 2x grayscale	GPS, IMU	7	43k	43k	[-80, +80]
ApolloScape [58]	2x	-	2x RGB, 2x grayscale	GPS, IMU	8	143k	20k	[-100, +100]
A2D2 [22]	5x 16-beam	-	6x	GPS, IMU	13	392k	40k	[-100, +100]
H3D [49]	1x 64-beam	-	3x	GPS, IMU	6	27k	27k	[-100, +100]
Cityscapes 3D [26]	-	-	1x Stereo Pair	-	2	25k	25k	[0, +200]
Lyft L5 [27]	1x	-	7x	-	8	170k	30k	[-100, +100]
A*3D [50]	1x 64-beam	-	1x Stereo Pair	-	3	39k	39k	[-100, +100]
SeeingThoughFog [5]	1x 64-beam, 1x 32-beam	1x	1x Stereo Pair, 1x Gated, 1x FIR	GPS,IMU	8	13.5k	13.5k	[-120, +120]
NuScenes [6]	1x 32-beam	5x	6x	GPS, IMU	14	400k	40k	[-100, +100]
NuPlan [7]	2x 20-beam, 3x 40-beam	-	8x	GPS, IMU	15	62.5M	(4.3M*)	[-100, +100]
NuImages [6]	-	-	1x out of 6	-	6	93k	93k	-
Waymo - Perception [53]	1x mid-range, 4x short-range	-	5x	-	10	230k	230k	[-100, +100]
Waymo - End2End [62]	-	-	8x	-	8	321k	-	-
ONCE [41]	1x 40-beam	-	7x	-	8	1M	16k	[-100, +100]
AiMotive [42]	1x 64-beam	2x	2x RGB, 2x RGB Fisheye	GPS, IMU	7	26.5k	26.5k	[-200, +200]
Argoverse V2 [60]	2x 32-beam	-	1x Stereo Pair, 7x Ring Cameras	GPS	11	150k	150k	[-250, +250]
MAN TruckScenes [20]	6x	6x	4x	GPS, IMU	18	30k	30k	[-226, +226]
TruckDrive (Ours)	7x long-range, 3x short-range	10x	1x / 3x Stereo Pair, 9x single	GPS, IMU	37	475k	165k	[-400, + 400]

mented by nuImages [6], a large-scale dataset focused on 2D object detection (OD), and nuPlan [7], the first large-scale, real-world benchmark for motion planning. Similarly, the Waymo Open Dataset ecosystem [53] offered an unprecedented scale. While initially focused on perception tasks, it has since expanded with the Waymo Motion dataset for trajectory forecasting [17] and the Waymo E2E benchmark for evaluating end-to-end driving models. The Argoverse datasets [10, 60] extended the common perception range up to 150 meters and the Lyft Level 5 dataset [27] focused on providing large-scale HD maps.

Task-Focused Autonomus Driving Datasets. Along with the development of large-scale benchmarks, several datasets have made significant contributions by focusing on specific tasks and modalities. Cityscapes 3D [26] extended the popular semantic segmentation benchmark [13, 14] with 3D bounding box annotations, bridging the gap between 2D and 3D scene understanding. ApolloScape [58] introduced a massive collection of data with a wide variety of tasks, including 3D detection, lane segmentation, and dense trajectory information for simulation. Datasets from automotive OEMs, such as A2D2 (Audi) [22], H3D (Honda) [49] and surround-view truck scenes from MAN Truckscenes [20], provide data from high-quality, industry-grade sensor configurations. KAIST dataset[35] and aiMotive [42] explored highway driving scenarios, although containing respectively only 1.2k annotated frames and 12k *highway* frames. The ONCE dataset [41] has pushed towards reducing annotation dependency by providing a large-scale benchmark for self-supervised learning, while A*3D [50] and SeeingThroughFog [5] explored active learning strategies or novel sensor setups to improve annotation efficiency in highly challenging weather conditions.

Limits of Existing Autonomous Driving Datasets. As presented in Table 1, prior autonomous driving datasets are dominated by urban, low speed setting and short effective ranges: 3D labels are rarely present above 80 meters, annotation density decrease rapidly with distance and long-range sensing is either absent or poorly represented [20, 60]. Moreover, they often offer low annotation amount and sensor modalities are limited to a small set of cameras and short range LiDARs, pushing models to fit specific biases and leaving safe heavy-vehicle driving as an open challenge.

3. TruckDrive Dataset

We introduce *TruckDrive*, a long-range, highway-focused dataset designed for heavy-vehicle autonomy. In this section, we first describe the TruckDrive domain and data collection process, emphasizing its diverse driving conditions, specialized sensor suite for high-speed perception and our cross-modal synchronization strategy. We further detail our annotation pipeline, which combines manual labeling with automated multi-view completion and kinematic refinement. Finally, we provide a quantitative analysis and comparison to foundational datasets (from Table 1), demonstrating gains in range, speed and trajectory coverage.

3.1. Dataset Domain

TruckDrive targets the driving domain of semi-trucks and other large commercial vehicles, covering scenarios that differ significantly from the urban, car-centric datasets commonly used in autonomous driving research. The dataset contains $3,828$ sequences recorded over 2 years across 8

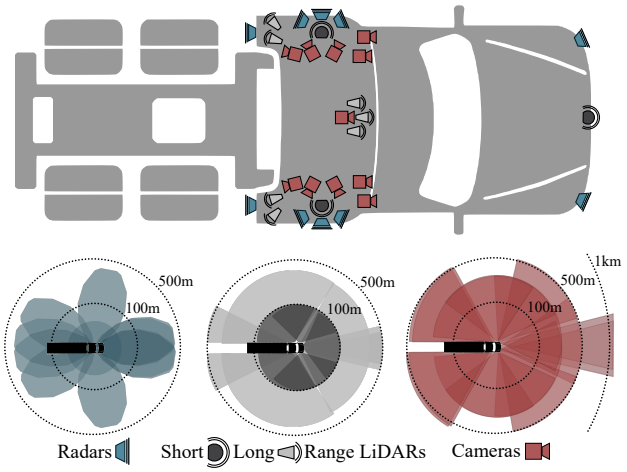


Figure 3. **Sensors Position and FoV.** Sensor position (top) and the nominal instrumented horizontal field of view (bottom) of, from left to right, radars, LiDARs and cameras, highlighting the unprecedented ranges at which they can operate.

Table 2. **Sensor Specifications and Raw Data Scale.** We present in detail our sensor platform, including RCCB cameras 3D short-range (SR) LiDARs, a 4D long-range (LR) FMCW LiDAR, and 4D radars, capturing 475 thousands synchronized frames

	Camera	**LiDAR**		**Radar**
	RCCB *AR0820*	**4D LR** *Aeries II*	**3D SR** *OS0/OS1*	**4D** *ARS540*
Make	OnSemi	AEVA	Ouster	Continental
Type	RCCB	FMCW 4D	3D	FMCW 4D
Resolution	3848×2168	$\sim$100 lines	$64/128 \times 2048$	—
FOV (H×V)	$52.8° \times 28.9°$	$120° \times 30°$	$360° \times 90°/45°$	$\pm 4°$–$\pm 20°$
f **(Hz)**	5–10	10		20
Raw Captures	6.3M	7.8M		6.0M
Sync Timestamps	569k	744k		601k
Cross-Modal Sync Timestamps: 475k				

U.S. states (NM, TX, VA, NC, TN, AR, WV, AZ), reaching a diversity-area metric [53] of $1,261.3$ km^2 ($16.5\times$ WOD). Data collection spans all seasons (48% fall, 32% winter, 15% spring, 5% summer) and diverse weather (80% sunny/cloudy/overcast, 10% fog, 10% precipitation). Sequences last 15–25 s with an average ego trajectory of 500 m, comprising mainly highways ($3,244$), followed by extra-urban (351) and urban roads (233). Driving patterns include 45.8% cruise/accelerate/brake, 36.5% lane changes/overtakes, 5.4% close cut-ins, and 12.3% complex layouts (work zones, intersections, unprotected turns). Illumination coverage includes $3,285$ daytime, 367 night, 122 dusk, and 54 dawn sequences.

3.2. Long-Range Sensor Setup

Our sensor suite, mounted on a semi-truck, is optimized for reliable perception in high-speed environments. Specifically, we employ 7 FMCW LiDARs (AEVA Aeries II), capable of measuring up to 400 meters and providing radial velocity, 3 short-range LiDARs (Ouster OS0/OS1), to account for blind spots and objects very close to the ego and 10 4D radars (Conti ARS540). Additionally, 11 to 15, depending on the configuration, RCCB cameras (9 short/medium focal and 1 to 3 long-focal stereo) provide high resolution imaging ($8MP$) at all ranges: QA verifies extrinsic accuracy below $0.015°$, bounding re-projection error beyond 200m. We report placement and horizontal coverage in Figure 3 and per-sensor specifications in Table 2.

FMCW Velocity. We rely on Frequency-Modulated Continuous-Wave (FMCW) technology, which allows to capture instantaneous radial velocity v_r for each point in the point cloud. The velocity measurement is derived from the Doppler-induced phase shift $\Delta\phi$ through

$$v_r = \frac{\Delta\phi \cdot \lambda}{4\pi} \cos\theta, \tag{1}$$

where λ is the wavelength and θ the angle of incidence.

Geo-Inertial Poses (PPK). For accurate ego motion we fuse data from 2 GNSS and 4 IMUs in a tailored Post-Processing Kinematic (PPK) pipeline, yielding reliable global poses for synchronized frames. Rare failure cases are complemented with LiDAR SLAM [48], providing ground-truth trajectories suitable for precise localization.

Sensors Synchronization Each different sensor group is triggered and synced to a common clock, allowing no more than 5 milliseconds between each unit capture. Cross-modal triggers are temporally aligned to enable near-simultaneous captures. Because our high-resolution cameras use a rolling shutter, showing a row-wise readout, aligning the other modalities to the image start time would induce a systematic temporal offset across rows. Instead, we define the reference timestamp at the image mid-exposure and synchronize LiDAR to this anchor

$$t_{\text{ref}} = t_{\text{img}}^{\text{start}} + \tfrac{1}{2} T_{\text{readout}}, \quad \left| t_{\text{LiDAR}} - t_{\text{ref}} \right| \leq 5\,\text{ms}, \tag{2}$$

with a typical T_{readout} of 54 milliseconds.

3.3. Annotation

We annotate 3D cuboids through a three-stage pipeline that combines human annotation with automated label refinement. To maximize the richness of the annotated data, human annotators manually curate sequential frames containing complex interactions or edge cases; in total, more than 2000 scenes are selected. Annotators then label 3D cuboids and 2D boxes and assign semantic classes. The selected annotations are subsequently refined automatically to enforce geometric and temporal consistency. For supervised learning tasks, the dataset provides around 140 k annotated training samples and 25 k annotated validation samples.

Stage 1: Human Annotation Primitives. During this stage, annotators produce geometric primitives consisting of 3D cuboids and 2D boxes (with relative Occlusion and

Truncation parameters) and assign semantic labels to all identified objects. 3D boxes are then iteratively adjusted using their projection into the cameras to reduce offset and avoid "ghost" objects. The annotation procedure results in 85 classes which we regroup in 9 main categories as shown in Figure 4a. The 9 classes to be captured are *traffic signs*, *passenger cars*, all types of road debris and interferences such as lost cargo, potholes, and cones (collectively referred to as *road obstructions*), *humans*, *semi-trucks* in both their cabins and trailers, 2-*wheeled vehicles*, *emergency vehicles* like police cars, ambulances or road-construction vehicles that can halter the nominal planning behavior and *vehicles* of different sizes, from heavy-duty vehicles, buses or single unit trucks to RV, trailers and equipment. Vulnerable Road Users are identified and included in the coarser categories.

Stage 2: Primitive Augmentation. For each timestamp we project the initial $3D$ cuboids into all camera views and match them against detections from a $2D$ object detector, by solving a bipartite assignment (Hungarian algorithm) with Intersection-over-Union as cost matrix. When a $2D$ detection has no correspondence, we fall back to the geometric projection or an existing $2D$ label. We handle truncations and perform class-wise Non Maximum Suppression (NMS) to promote high confidence $2D$ detections, resulting in the set of matched 3D detections and 2D-only candidates.

Stage 3: Refinement and Completion. The existing matched $3D$ annotations are transformed into a global coordinate frame and their trajectories are refined through a kinematically constrained optimization, enforcing plausible motion and reducing yaw jitter. Specifically we minimize

$$\min_{\{s_t^k, d_t^k\}} \sum_{t \in \mathcal{T}_k} \left[\lambda_{\text{o}} L_t^{\text{o}} + \lambda_\psi L_t^\psi + \lambda_d L_t^d + \lambda_{\text{smooth}} L_t^{\text{sm}}\right], \quad (3)$$

$$L_t^{\text{o}} = \rho(\|c(s_t^k) - \hat{c}_t^k\|_2), \quad L_t^\psi = \rho(\text{ang}(\psi_t^k, \hat{\psi}_t^k)), \quad (4)$$

$$L_t^d = \rho(\|d_t^k - \hat{d}_t^k\|_2), \quad L_t^{\text{sm}} = \|\Delta v_t^k\|_2^2 + \|\Delta^2 \psi_t^k\|_2^2. \quad (5)$$

subject to a unicycle model

$$\begin{aligned} x_{t+1}^k &= x_t^k + \Delta t\, v_t^k \cos\psi_t^k, \quad & \psi_{t+1}^k &= \psi_t^k + \Delta t\, \omega_t^k, \\ y_{t+1}^k &= y_t^k + \Delta t\, v_t^k \sin\psi_t^k, & \kappa_t^k &= \omega_t^k / v_t^k. \end{aligned} \quad (6)$$

Here, $s_t^k = (x_t^k, y_t^k, \psi_t^k, v_t^k, \omega_t^k)$ is the per-track state, $d_t^k = (\ell_t^k, w_t^k, h_t^k)$ are box sizes, $c(s_t^k) = (x_t^k, y_t^k)$ extracts the box center, hats $\hat{\cdot}$ denote noisy estimates, $\rho(\cdot)$ is a robust loss (Huber with scale δ_ρ), ang is the angle difference, Δ/Δ^2 are first and second finite differences.

For short gaps $t \in [t_1, t_2]$ with missing frames we initialize bounding boxes by interpolating

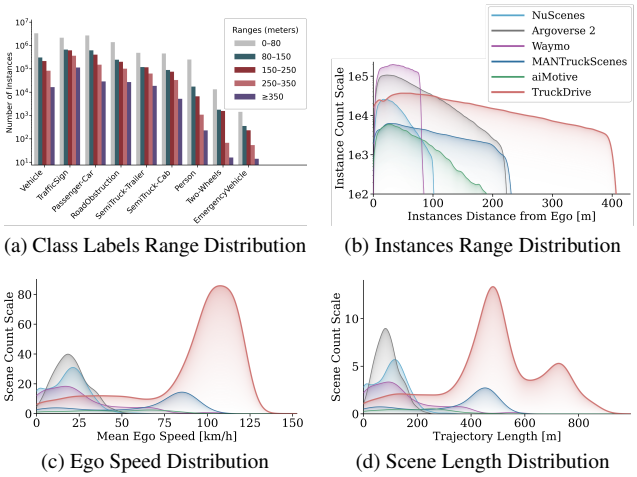
$$\begin{aligned} \tilde{c}_t &= (1-\alpha)c_{t_1} + \alpha c_{t_2}, \quad & \tilde{\psi}_t &= \text{slerp}(\psi_{t_1}, \psi_{t_2}; \alpha), \\ \tilde{d}_t &= (1-\alpha)d_{t_1} + \alpha d_{t_2}, & \alpha &= (t-t_1)/(t_2-t_1), \end{aligned} \quad (7)$$

then refine jointly using Equations (3) to (6).

Concurrently, we lift unmatched $2D$ candidates from Stage 2 into $3D$. For each camera c, we project the eight cuboid corners of a 3D hypothesis $p = (x, y, z, \ell, w, h, \psi)$ and form the tight axis-aligned 2D box $\hat{b}_c(p)$. We retain only those camera views whose Stage 2 detection $b_{c,t} = [x_0, y_0, x_1, y_1]$ has sufficient overlap with the hypothesis, defined as $\text{IoU}\big(\hat{b}_c(p),\, b_{c,t}\big) \geq 0.3$, and optimize p so that the projected boxes fit the detections across the retained views

$$\sum_{c \in \mathcal{C}} \Big[\lambda_{\text{iou}}\big(1 - \text{IoU}(\hat{b}_c(p), b_c)\big) + \lambda_g\big(z_{\min}(p) - z_g\big)^2\Big], \quad (8)$$

where z_g is the local ground height from the accumulated LiDAR map. 3D objects are then tracked over time with a offline tracker [59], identity-aligned to ground truth via temporal IoU voting and merged with the smoothed ground-truth boxes to form the final annotation set.



(a) Class Labels Range Distribution (b) Instances Range Distribution

(c) Ego Speed Distribution (d) Scene Length Distribution

Figure 4. **Dataset Analysis.** Our dataset comprises an unprecedented density of instance objects at ranges (greater than 200 meters) yet to be explored in publicly available datasets (a,b), as well as driving speeds 5 times higher (c) and sequences with traveled length up to 8 times longer (d) than existing benchmarks.

3.4. Dataset Analysis

TruckDrive, compared in Table 1 with other benchmarks, introduces an unprecedented sensing configuration with 37 heterogeneous sensors, double the number available in the second most sensor-rich dataset (18), enabling full 360° perception coverage with both long and short-range redundancy and enhancing robustness in complex environments. TruckDrive's LiDAR extends up to 400 meters in both the forward and rear directions, twice the maximum range reported in previous benchmarks (220 m). The dataset comprises approximately 165,000 manually annotated frames, which is comparable in scale to the largest publicly available datasets (230 k). Per-class instances are distributed uniformly across the full perception range, yielding balanced near and far-field samples (Fig. 4a). The density of annotated 3D boxes decays gradually with distance up to 400 m (Fig. 4b), while 2D boxes extend well beyond 1000 m, in contrast to prior urban-focused datasets [6, 53, 60] where

annotations beyond $100 - 200$ m are rare and instance density drops sharply after 80 m. Figures 4c and 4d highlight highway dynamics in TruckDrive. Speeds span from low on/off ramp segments to up to 130 km/h, surpassing urban datasets capped below 75 km/h. Sequences extend to 900 m (against 400 m of urban datasets), enabling temporal reasoning at high speed and more faithful evaluation of long-horizon perception and prediction.

4. Driving Tasks and Challenges

We use the proposed dataset at hand to evaluate recent perception and driving methods across typical tasks, such as 2D and 3D object detection, tracking, depth estimation, LiDAR forecasting, moving object segmentation, 3D scene reconstruction and end-to-end planning. This evaluation investigates whether current state-of-the-art approaches, primarily developed and optimized for urban driving datasets, can generalize to the speed, long-range and large-scale highway scenarios present in TruckDrive. To this end, all tested models have been trained on our TruckDrive data. We train all models with a consistent train-validation split made of 140 and 25 thousand samples respectively and follow standard metrics and protocols. We couple quantitative with qualitative results for the target domain in Figure 5.

4.1. 2D Object Detection

In nuScenes and KITTI [6, 21], 2D performance is largely driven by 3D detectors due to low image resolution and wide FOV; 3D NMS in lifted space handles occlusion better than image-space NMS. At kilometer ranges, however, objects in those benchmarks would be sub-pixel, whereas our 8MP imagery keeps them resolvable, so only 2D detectors are able detect them. We train state-of-the-art architectures [8, 32, 36, 66] and report results in Table 3.

4.2. 3D Object Detection

We evaluate long-range 3D object detection using three SOTA models on our dataset, spanning a LiDAR based model [1], a camera-based method [31] and a common LiDAR-camera fusion architecture [40]. We report average precision over three range bins in Table 4.

4.3. 3D Multi Object Tracking

We evaluate whether state-of-the-art tracking methods can handle the long-horizon scenes and high differential velocities between the ego and other agents in TruckDrive, which stress association over long gaps and occlusions. We report MOT results for a query based approach [67] and two 3D boxes based methods [59, 65] in Table 5.

4.4. Depth Estimation

We train monocular, stereo and surround depth estimation models under long-range LiDAR supervision to assess the capability of current approaches in the TruckDrive domain. For all subtasks, we report standard task metrics alongside unified, distance-binned depth metrics, ensuring balanced evaluation across ranges and avoiding the near-range bias and limited range-dependent interpretability of disparity-based or relative-error metrics.

Table 3. **2D Object Detection Results.** We follow CoCo [39] and report mean average precision (mAP) at 0.50 IoU, mAP at 0.75 IoU, and mAP at short ($0 - 50$ m, SR), medium ($50 - 150$ m, MR), long ($150 - 250$ m, LR), and ultra-long-range (250+, UR).

Method	mAP↑	mAP_{50}↑	mAP_{75}↑	mAP_{SR}↑	mAP_{MR}↑	mAP_{LR}↑	mAP_{UR}↑
DETR [8]	12.70%	23.90%	12.20%	41.20%	24.70%	8.90%	1.00%
ViTDet [36]	27.30%	37.60%	30.80%	58.30%	51.80%	33.90%	3.30%
YOLO11x [32]	28.90%	39.00%	31.60%	36.30%	29.40%	8.20%	2.00%
DINO [66]	37.80%	54.20%	40.30%	63.90%	54.60%	43.20%	15.30%

Table 4. **3D Object Detection Results**. We report mAP for 3 baselines using a single or a combination of LiDAR (L) and Camera (C) data, divided into short ($0 - 50$ m, SR), medium ($50 - 150$ m, MR), long ($150 - 250$ m, LR) and full detection ranges.

Method	Mode	mAP↑ Full	mAP↑ SR	mAP↑ MR	mAP↑ LR
Far3D [31]	C	14.04%	35.54%	11.07%	0.33%
TransFusion-L [1]	L	25.24%	30.12%	22.25%	22.25%
BEVFusion [40]	L+C	26.45%	32.32%	22.77%	22.69%

Table 5. **3D Multi Object Tracking Results**. We report AMOTA, AMOTP and Recall for a query based and two LiDAR based methods.† uses [40] inference detections.

Method	Mode	AMOTA↑	AMOTP↓	Recall↑
MUTR3D [67]	Query	6.1%	79.0%	11.4%
Immortal Tracker † [59]	3D Box	12.8%	77.2%	20.7%
CenterPoint † [65]	3D Box	13.0%	76.9%	21.5%

Depth Evaluation Ground-Truth. For our benchmark, we build dense LiDAR ground truth by accumulating static points and filtering dynamic objects through the FMCW capabilities of our 4D LiDAR. The resulting depth map is projected into each frame, where we reintroduce dynamic points based on their timestamps, filter out view-dependent occlusions and enhance temporal consistency using dense depth priors inferred from an ensemble of depth foundation models. Additional details in the Supplementary Material.

Surround Views. Leveraging the wide, calibrated overlap among five high-resolution cameras arranged to ensure extensive, overlapping surround coverage, we train two state-of-the-art models [33, 52] for metric surround depth estimation and report results in Table 6a, evaluated against the dense LiDAR ground truth. Task-specific relative metrics are reported following [52].

Stereo Views. The forward-facing cameras are arranged in a wide-baseline stereo configuration (approx. 1.57 m), providing a strong geometric basis for depth perception via triangulation. We evaluate state-of-the-art learning-based stereo matching methods [12, 24, 25] and report results in

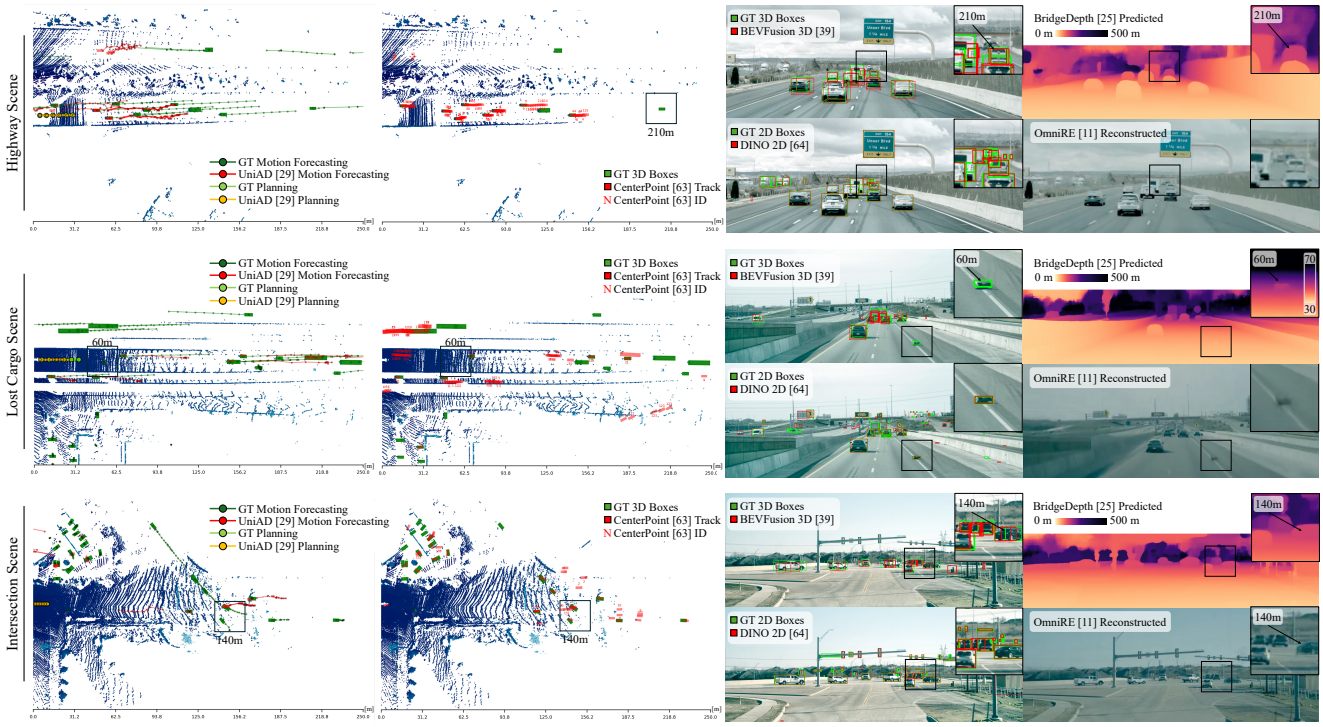


Figure 5. **Driving Tasks and Challenges.** We report qualitative results of the best baselines across planning, 2D/3D object detection, depth estimation and scene reconstruction. Even when trained on TruckDrive, existing methods struggle in the long-range, high-speed regime. Planning modules exhibit conservative behavior due to low-speed assumptions. Grid-based BEV models degrade perception as large spatial coverage demands heavy downsampling, erasing safety-critical details such as small debris or lost cargo [40], while depth methods struggle beyond 200m and in sky regions [25], revealing limited distance awareness and motivating architectures for highway-scale perception.

Table 6b. We report task-specific disparity metrics following the KITTI stereo benchmark [43, 44].

Monocular View. We benchmark recent existing monocular depth estimation models [4, 28, 51], which infer depth from single images without geometric priors, to assess their ability to generalize to the scale and appearance of distant objects. Each model is trained twice: once using the same 5 cameras employed for surround views, and once using the left stereo camera, enabling direct comparison with stereo and surround-view architectures. Results are reported in Table 6c. Task-specific metrics are reported following the KITTI benchmark [56] for monocular depth estimation.

4.5. Temporal Scene Modeling and Reconstruction

Predicting future scene geometry is fundamental for safe motion planning. We benchmark recent methods on the LiDAR forecasting task over a challenging 250 meters Region Of Interest (ROI) ahead of the ego vehicle, comparing a LiDAR-only [34], a camera-only [64], and a multi-modal fusion network [47]. We report range-binned results in Table 7. For dynamic modeling, we evaluate a LiDAR-based moving-object segmentation method [45] chosen for its strong out-of-domain generalization. As shown in Table 8, the pretrained model struggles at longer distances, indicating the need for long-range training to improve detection. Beyond discrete object-level tasks, high-fidelity scene

Table 6. **Depth Estimation Results**. We report performances for surround (a), stereo (b) and monocular (c) depth estimation tasks. Each method is evaluated with standard accuracy and error metrics at short (0-50m, SR), medium (50-150m, MR), long (150-250m, LR) and ultra (250-1000m, UR) range bins.

(a) Multi-Camera Surround Depth Estimation

	Distance-Binned MAE (Depth)				Task-Specific Depth Metrics			
Method	**SR ↓**	**MR ↓**	**LR ↓**	**UR ↓**	**Abs Rel ↓**	**Sq Rel ↓**	**RMSE ↓**	**δ_1 ↑**
R3D3 [52]	7.99	25.35	74.02	181.22	0.30	0.21	37.60	0.49
MapAnything [33]	5.05	16.73	39.19	121.15	0.19	6.13	26.40	0.73

(b) Stereo Disparity Estimation

	Distance-Binned MAE (Depth)				Task-Specific Disparity Metrics		
Method	**SR ↓**	**MR ↓**	**LR ↓**	**UR ↓**	**D1-bg ↓**	**D1-fg ↓**	**D1-all ↓**
NMRF [24]	3.39	9.13	20.92	40.88	26.98	21.95	21.95
MonSter++ [12]	4.41	9.21	21.39	62.18	26.09	23.94	29.07
BridgeDepth [25]	2.53	8.34	20.21	69.10	28.74	11.12	28.57

(c) Monocular Depth Estimation

	Distance-Binned MAE (depth)				Task-Specific Depth Metrics			
Method	**SR ↓**	**MR ↓**	**LR ↓**	**UR ↓**	**SILog↓**	**sqERel↓**	**absERel↓**	**iRMSE↓**
Multi-View								
ZoeDepth [4]	4.77	17.63	44.00	114.30	27.25	0.13	0.16	8.54
Metric3Dv2 [28]	4.68	15.26	42.11	144.51	25.31	0.11	0.17	9.06
UniDepthv2 [51]	3.52	12.30	28.63	103.94	21.07	0.06	0.14	2.85
Single-View								
ZoeDepth [4]	4.15	15.80	45.55	133.78	23.93	0.07	0.20	3.51
Metric3Dv2 [28]	3.28	12.81	27.53	94.47	22.01	0.05	0.14	2.75
UniDepthv2 [51]	2.66	10.63	28.37	102.58	20.08	0.05	0.13	2.45

Table 7. **LiDAR Forecasting Results**. We evaluate single and multi-modal state of the arts methods with Chamfer Distances (CD) of 1 and 3 seconds and L1 error.

History Horizon	Method	Modality	1s		3s	
			CD ↓	L1 (m) ↓	CD ↓	L1 (m) ↓
1s	4DOcc [34]	L	18.93	4.69	-	-
	ViDAR [64]	C	58.23	20.21	51.72	20.69
	LRS4Fusion [47]	L + C	15.82	3.31	39.03	3.82
3s	4DOcc [34]	L	23.58	3.00	47.81	4.29
	ViDAR [64]	C	57.28	20.14	56.20	20.53
	LRS4Fusion [47]	L + C	16.38	2.49	42.93	4.05

Table 8. **3D Moving Object Segmentation Results**. ‡ indicates results from the public KITTI [21] checkpoint.

Method	SR		MR		LR		FULL	
	F1↑	IoU ↑	F1↑	IoU ↑	F1↑	IoU ↑	F1↑	IoU ↑
4DMOS‡ [45]	25.9	18.5	8.4	6.1	0.6	0.4	24.4	16.7
4DMOS [45]	47.3	32.1	22.7	15.4	8.3	5.6	31.8	21.6

Table 9. **3D Reconstruction Quality Results.** We report PSNR and SSIM for a NeRF and two 3D Gaussian Splatting methods.

Method	Representation	PSNR ↑	SSIM ↑
Dyn. Nerfacto [55]	NeRF	26.2870	0.8653
HUGS [70]	3DGS	29.2675	0.8858
OmniRe [11]	3DGS	33.8244	0.9515

Table 10. **E2E Planning.** We train UniAD [29] on our long range setup and evaluate L2 error for all predicted time intervals.

Method	$L2$ (m)↓						
	1 step	2 step	3 step	4 step	5 step	6 step	Avg.
UniAD [29]	0.57	1.13	1.71	2.30	2.88	3.42	2.00

reconstruction on long-range data is critical for photorealistic digital twins and dense scene understanding. Therefore, we assess a Neural Radiance Fields (NeRF) [55] and two 3D Gaussian Splatting (3DGS) methods [11, 70] in Table 9.

4.6. End2End Driving

Collectively, all tasks above aim at enabling end-to-end planning aligned with TruckDrive's goal of safe, reliable and proactive operation. We train and evaluate UniAD [29] as a recent E2E driving method, extending the ROI from 50 m to 250 m and replacing the original camera-only [37] BEV backbone with a LiDAR-based architecture [1], offering a first E2E benchmark for long-range highway driving. We evaluate UniAD on open-loop planning with standard L2 error, see results in table 10.

5. Discussion

Our experiments confirm that across all tasks, existing model architectures designed for publicly available short-range data underperform when trained on TruckDrive's long-range regime, with scores monotonically dropping with distance. Camera-only models exhibit the lowest performance, with average 57% lower mAP for 2D object detection and up to 99% lower mAP for 3D object detection (Far3D [31]) in far (LR) distances. Architectures relying on camera, limited by compute constraints, necessitate 3× downsampling of native 8MP inputs, substantially degrading performance; for instance, Long Range stereo depth estimation exhibits an 8× MAE increase (BridgeDepth [25]) due to reduced pixel disparities. LiDAR based and fusion-based architectures are aided in training by the additional long range 3D representation, but struggle in sustaining the high dimensional complexity of the data and the large translation of objects. As existing methods largely rely on dense BEV representations, extending the maximum range forces either larger grids with fixed resolution, inducing a quadratic memory growth, or coarser cells with fixed grid dimension, degrading localization and association of both smaller objects and far-range instances, as shown Figure 5. As a result, 3D multi-object tracking performs poorly (average 10% AMOTA), and we observe drops up to 83% for moving-object segmentation (4DMOS [45]) and up to 31% for long-range 3D object detection (BEVFusion [40]). Finally, UniAD [29] requires extensive down-sampling across the entire architecture to allow the model to fit in the memory. The 250 × 250 meters ROI is encoded in a 200 × 200 BEV grid over the entire implementation, too coarse to encode useful driving information and not accurate enough to compute meaningful collision metric values. Overall, the model struggles to achieve low L2 planning error even for close future timestamps (3 step: 1.71 m), showcasing how urban-centric architectures fail to scale to long-range and high speed scenarios, highlighting the need for further research to unlock safe and reliable highway driving.

6. Conclusion

We introduce an autonomous driving dataset with 2D annotations up to 1 km and 3D annotations up to 400 m tailored for highway driving. While existing datasets focus on urban passenger car driving, the proposed TruckDrive dataset aims at opening up the research to highway driving where higher speed requires the ego agent to use different trajectories and maneuvers. We specifically focus on heavy-duty commercial trucks, which present an additional layer of complexity due the immense mass and break system lags extending the useful perception range from 80 m to 400 m.

Our evaluations on the dataset expose a persistent gap between state-of-the-art methods and the requirements of trucking highway autonomy. Hence, the dataset establishes a benchmark for range-aware, temporally grounded and computationally efficient driving methods that operate safely and reliably at high speed over long distances, and serves as a foundation for future research into driving methods tailored to the unique challenges of highway-scale autonomy, still far less explored than their urban counterpart.

7. Acknowledgments

Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award and a Amazon Science Research Award. Felix Heide is a co-founder of Algolux (now Torc Robotics), Head of AI at Torc Robotics, and a co-founder of Cephia AI.
We thank everyone at Torc Robotics for making TruckDrive possible, in particular the teams responsible for sensor integration, fleet operations, data engineering, and tooling for dataset generation and curation.

References

[1] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, 2022. 1, 6, 8

[2] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019. 2, 3

[3] Jens Behley, Andres Milioto, and Cyrill Stachniss. A benchmark for lidar-based panoptic segmentation based on kitti, 2020. 2

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 7

[5] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather, 2020. 3

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 5, 6

[7] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles, 2022. 3

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 6

[9] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository, 2015. 2

[10] Ming-Fang Chang, John W Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[11] Ziyu Chen, Jiawei Yang, Jiahui Huang, Riccardo de Lutio, Janick Martinez Esturo, Boris Ivanovic, Or Litany, Zan Gojcic, Sanja Fidler, Marco Pavone, Li Song, and Yue Wang. Omnire: Omni urban scene reconstruction. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[12] Junda Cheng, Wenjing Liao, Zhipeng Cai, Longliang Liu, Gangwei Xu, Xianqi Wang, Yuzhou Wang, Zikang Yuan, Yong Deng, Jinliang Zang, Yangyang Shi, Jinhui Tang, and Xin Yang. Monster++: Unified stereo matching, multi-view stereo, and real-time stereo with monodepth priors, 2025. 6, 7

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015. 3

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[16] Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, and Juergen Gall. 3dmotformer: Graph transformer for online 3d multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9784–9794, 2023. 1

[17] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9710–9719, 2021. 2, 3

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010. 2

[19] Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fsd v2: Improving fully sparse 3d object detection with virtual voxels, 2023. 1

[20] Felix Fent, Fabian Kuttenreich, Florian Ruch, Farija Rizwin, Stefan Juergens, Lorenz Lechermann, Christian Nissler, Andrea Perl, Ulrich Voll, Min Yan, and Markus Lienkamp. Man truckscenes: A multimodal dataset for autonomous trucking in diverse conditions, 2024. 3

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark

suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 1, 2, 3, 6, 8

[22] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2d2: Audi autonomous driving dataset, 2020. 3

[23] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2022. 1

[24] Tongfan Guan, Chen Wang, and Yun-Hui Liu. Neural markov random field for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2024. 1, 6, 7

[25] Tongfan Guan, Jiaxin Guo, Chen Wang, and Yun-Hui Liu. Bridgedepth: Bridging monocular and stereo reasoning with latent alignment. *To appear*, pages 27681–27691, 2025. 6, 7, 8

[26] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection, 2020. 3

[27] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset, 2020. 3

[28] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 7

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 8

[30] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *ICCV*, 2023. 1

[31] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection, 2023. 2, 6, 8

[32] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 6

[33] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *arXiv:2509.13414*, 2025. 6, 7

[34] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7, 8

[35] Byungju Kim, Junho Yim, and Junmo Kim. Highway driving dataset for semantic video segmentation, 2020. 3

[36] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 6

[37] 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 *European conference on computer vision*, pages 1–18. Springer, 2022. 1, 8

[38] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *Pattern Analysis and Machine Intelligence (PAMI)*, 2022. 2

[39] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2, 6

[40] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 1, 6, 7, 8

[41] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, Hang Xu, and Chunjing Xu. One million scenes for autonomous driving: Once dataset, 2021. 3

[42] Tamás Matuszka, Iván Barton, Ádám Butykai, Péter Hajas, Dávid Kiss, Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh, Levente Pető, Dezső Ribli, Dávid Szeghy, Szabolcs Vajna, and Bálint Varga. aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception. 2022. 3

[43] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 2, 7

[44] Moritz Menze, Christian Heipke, and Andreas Geiger. Object scene flow. *ISPRS Journal of Photogrammetry and Remote Sensing (JPRS)*, 2018. 2, 7

[45] B. Mersch, X. Chen, I. Vizzo, L. Nunes, J. Behley, and C. Stachniss. Receding moving object segmentation in 3d lidar data using sparse 4d convolutions. *IEEE Robotics and Automation Letters (RA-L)*, 7(3):7503–7510, 2022. 7, 8

[46] Arian Mousakhan, Sudhanshu Mittal, Silvio Galesso, Karim Farid, and Thomas Brox. Orbis: Overcoming challenges of long-horizon prediction in driving world models. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 2

[47] Edoardo Palladin, Samuel Brucker, Filippo Ghilotti, Praveen Narayanan, Mario Bijelic, and Felix Heide. Self-supervised

sparse sensor fusion for long range perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 2, 7, 8

[48] Yue Pan, Xingguang Zhong, Louis Wiesmann, Thorbjörn Posewsky, Jens Behley, and Cyrill Stachniss. Pin-slam: Lidar slam using a point-based implicit neural representation for achieving global map consistency. *IEEE Transactions on Robotics*, 40:4045–4064, 2024. 4

[49] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes, 2019. 3

[50] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A*3d dataset: Towards autonomous driving in challenging environments. In *Proc. of The International Conference in Robotics and Automation (ICRA)*, 2020. 3

[51] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: Universal monocular metric depth estimation made simpler, 2025. 1, 7

[52] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3216–3226, 2023. 6, 7

[53] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3, 4, 5

[54] Manuel Muñoz Sánchez, Chris van der Ploeg, Robin Smit, Jos Elfring, Emilia Silvas, and Rene van de Molengraft. Prediction horizon requirements for automated driving: Optimizing safety, comfort, and efficiency. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2575–2582, 2024. 2

[55] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 8

[56] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *International Conference on 3D Vision (3DV)*, 2017. 7

[57] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, 2018. Association for Computational Linguistics. 2

[58] Peng Wang, Xinyu Huang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[59] Qitai Wang, Yuntao Chen, Ziqi Pang, Naiyan Wang, and Zhaoxiang Zhang. Immortal tracker: Tracklet never dies. *arXiv preprint arXiv:2111.13672*, 2021. 1, 5, 6

[60] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021. 1, 3, 5

[61] Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, and Wei Zhan. Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection. *arXiv preprint arXiv:2304.14340*, 2023. 2

[62] Runsheng Xu, Hubert Lin, Wonseok Jeon, Hao Feng, Yuliang Zou, Liting Sun, John Gorman, Kate Tolstaya, Sarah Tang, Brandyn White, Ben Sapp, Mingxing Tan, Jyh-Jing Hwang, and Drago Anguelov. Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios, 2025. 3

[63] Harsh Yadav, Maximilian Schaefer, Kun Zhao, and Tobias Meisen. *CASPFormer: Trajectory Prediction from BEV Images with Deformable Attention*, page 420–434. Springer Nature Switzerland, 2024. 2

[64] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 7, 8

[65] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking, 2021. 1, 6

[66] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 6

[67] Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. *arXiv preprint arXiv:2205.00613*, 2022. 1, 6

[68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[69] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 2

[70] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages
21336–21345, 2024. 8