DurLAR: A High-Fidelity 128-Channel LiDAR Dataset with Panoramic Ambient and Reflectivity Imagery for Multi-Modal Autonomous Driving Applications

Li Li[1] Khalid N. Ismail[1] Hubert P. H. Shum[1] Toby P. Breckon[1,2]
Department of {Computer Science[1] | Engineering[2]}, Durham University, UK
{li.li4, khalid.n.ismail, hubert.shum, toby.breckon}@durham.ac.uk

Abstract

We present DurLAR, a high-fidelity 128-channel 3D LiDAR dataset with panoramic ambient (near infrared) and reflectivity imagery, as well as a sample benchmark task using depth estimation for autonomous driving applications. Our driving platform is equipped with a high resolution 128 channel LiDAR, a 2MPix stereo camera, a lux meter and a GNSS/INS system. Ambient and reflectivity images are made available along with the LiDAR point clouds to facilitate multi-modal use of concurrent ambient and reflectivity scene information. Leveraging DurLAR, with a resolution exceeding that of prior benchmarks, we consider the task of monocular depth estimation and use this increased availability of higher resolution, yet sparse ground truth scene depth information to propose a novel joint supervised/self-supervised loss formulation. We compare performance over both our new DurLAR dataset, the established KITTI benchmark and the Cityscapes dataset. Our evaluation shows our joint use supervised and self-supervised loss terms, enabled via the superior ground truth resolution and availability within DurLAR improves the quantitative and qualitative performance of leading contemporary monocular depth estimation approaches ($RMSE = 3.639$, $SqRel = 0.936$).

1. Introduction

LiDAR (Light Detection and Ranging) is one of the core perception technologies enabling future self-driving vehicles and advanced driver assistance systems (ADAS). Multiple datasets featuring LiDAR have been proposed to evaluate semantic in geometric scene understanding tasks such as depth estimation, object detection, visual odometry, optical flow and tracking [30, 49, 24, 11, 35, 34, 10, 44, 53, 52]. Based on this existing dataset provision, various architectures have been proposed for LiDAR based scene understanding in this domain [7, 9, 27, 59, 22, 20, 1, 8]. Moreover, benchmarks and evaluation metrics have emerged to facilitate the comparison of varies techniques and datasets [25, 55, 29, 5, 43].

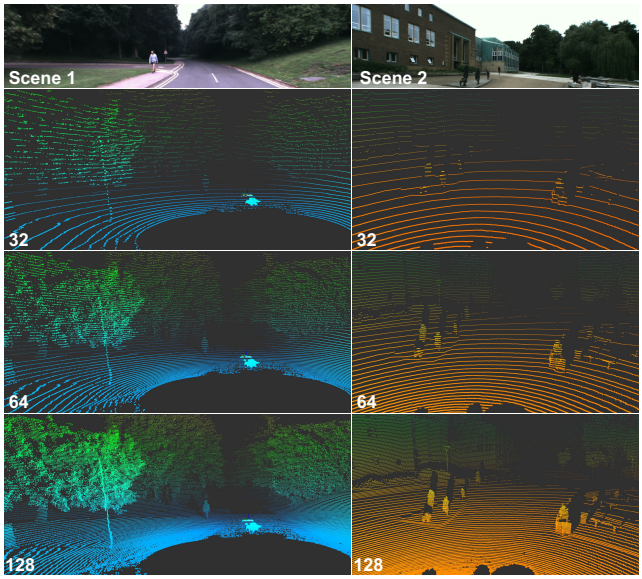

Figure 1: LiDAR point clouds from two exemplar scenes with differing vertical LiDAR resolution (top to bottom: colour RGB images, [32 → 64 → 128] LiDAR channels).

In these datasets, LiDAR range data corresponding to the colour image of the environment is provided as the ground-truth depth information. Such ground truth can be relatively sparse compared to the sampling of the corresponding colour camera imagery — typically as low as 16 to 64 channels of depth (see Figure 1, *e.g*, 16-64 horizontal scanlines of depth information, spanning 360 degrees from the vehicle over a 50-200 m range). Here, the terminology *channel* refers to the vertical resolution of the LiDAR scanner, and has a one-to-one correspondence to the laser beam as it is referred to in some studies. With this in mind, current datasets and their associated metric-driven benchmarks are significantly limited when compared to the contemporary availability of high-resolution LiDAR data as we pursue in this paper.

By contrast, we propose a large-scale high-fidelity LiDAR dataset[1] based on the use of a 128 channel LiDAR unit mounted on our Renault Twizy test vehicle (Figure 2). Compared to existing LiDAR datasets in this field (Table 1), including the seminal KITTI dataset [24, 26, 46], our dataset has the following novel features:

- **High vertical resolution LiDAR**, which offers both superior spatial depth resolution (Figure 1) and additionally co-registered 360°ambient and reflectivity imagery that is concurrently captured via the LiDAR laser return itself.
- **Additional synchronised sensors** including a high resolution forward-facing stereo imagery (2MPix), a high fidelity GNSS/INS and a lux meter.
- **Route repetition** such that the dataset uses the same set of driving routes under varying environmental conditions, such as overcast, rainy weather, seasonal variations and varying times of day - hence facilitating evaluation under different weather and illumination conditions.

Subsequently, our dataset is presented as a KITTI-compatible offering such that the data formats used can be parsed using both our DurLAR development kit and the official KITTI tools (in addition to third party KITTI tools).

Figure 2: **Test vehicle** (Renault Twizy): equipped with a long range stereo camera, a LiDAR, a lux meter and a combined GNSS/INS inertial navigation system.

In order to illustrate the advantages and potential applications of this proposed benchmark dataset, we adopt monocular depth estimation as a sample task for comparison. We thus evaluate the relative performance of contemporary monocular depth estimation architectures [27, 60, 61], by leveraging the higher resolution LiDAR capability within DurLAR to facilitate more effective use of depth supervision, for which we propose a novel joint supervised/self-supervised loss formulation (Section 4).

More broadly, the illumination-independent sensing capabilities of high-resolution 3D LiDAR additionally enable the evaluation of a range of driving tasks [54, 11] under varying environmental conditions spanning both extreme weather and illumination changes using our dataset.

Our main contributions are summarised as follows:

- a novel large-scale dataset comprising contemporary high-fidelity 3D LiDAR (128 channels), stereo/ambient/reflectivity imagery, GNSS/INS and environmental illumination information under repeated route, variable environment conditions (in the *de facto* KITTI dataset format). The first autonomous driving task dataset to additionally comprise usable ambience and reflectivity LiDAR obtained imagery (360°, 2048×128 resolution).
- an exemplar monocular depth estimation benchmark to compare the performance of supervised/self-supervised variants of three leading approaches [70, 27, 61] when trained and evaluated on low resolution (KITTI [24]), high resolution (DurLAR) ground truth LiDAR depth data, or our novel KITTI/DurLAR dataset partition, with the observation that increased resolution and availability enables superior monocular depth estimation performance via the use of our joint supervised/self-supervised loss formulation (Table 3, Table 4, Figure 9).

2. Related Work

We consider prior work in two related topic areas: autonomous driving datasets (Section 2.1) and monocular depth estimation (Section 2.2).

2.1. Autonomous Driving Datasets

There are multiple autonomous driving task datasets that provide 3D LiDAR data for outdoor environments (Table 1).

High vertical resolution LiDAR is not present in existing datasets (see Table 1). The vertical resolution of LiViSet [11] and nuScenes [10] is 32 channels. Similarly the Stanford Track Collection [53], KITTI [24], Sydney Urban Objects [52], DENSE [30], H3D [49], SemanticKITTI [5], Lyft Level 5 [35, 34] and KITTI-360 [41] is 64 channels. In contrast, our proposed dataset has a higher vertical resolution of 128 channels, which can capture a significantly higher level of detail of environment objects (Figure 1).

Rolling shutter effect is common among analogue spinning LiDAR, *e.g.*, Velodyne scanners, which are widely used in most of the existing datasets [24, 52, 11, 10, 30, 49, 5, 35, 41]. Instead, the Ouster LiDAR we use is a multi-beam flash LiDAR [47], meaning all 128 channels are shot simultaneously, avoiding this distortion effect.

In adverse weather, LiDAR fails [51] (*e.g.* fog), since opaque particles will distort light and reduce visibility significantly, whilst it can produce fine-grained point clouds

[1] Online access for the dataset, https://github.com/l1997i/DurLAR.

Dataset	Resolution	Range/m	Diversity	Image	# Frames	Other sensors
DENSE [30]	64	120	E/W/T	I	1M	D/M/F/T/B
H3D [49]	64	120	E	I	28k	G/M
KITTI | SemanticKITTI | KITTI-360 [24, 5, 41]	64	120	E	I	93k|93k|320k	N/S/G/M/B
LiVi-Set [11]	32	100	E	I	10k	
Lyft Level 5 [35, 34]	64	200	E/W/T	I	170k	D/B
nuScenes [10]	32	100	E/W/T	I	1M	M/D/B
Oxford RobotCar [44]	4[a]	50	E/W/T	I	3M[b]	N/S/G/M/B
Stanford Track Collection [53]	64	120	E	I	14k	M
Sydney Urban Objects [52]	64	120	E	I	0.6k[c]	
DurLAR (ours)	**128**	**120**	**E/W/T/L**	**I/A/R**	**100k**	**U/N/S/G/M/B**

Table 1: Existing public LiDAR datasets for autonomous driving tasks detailing vertical resolution (# channels), diversity in terms of environments (E), times of day (T), weather conditions (W), same route of repeated locations (L) and also the type of LiDAR images made available in addition to range information as: intensity (I), ambient (A), reflectivity (R). Other sensors refer to radar (D), lux meter (U), GNSS supporting more than 2 constellations (N), INS (S), GPS (G), IMU (M), FIR camera (F), Near infrared camera (T) and stereo camera (B). [a] the number of planes. SICK LD-MRS LiDAR has 4 planes, and SICK LMS-151 LiDAR has 1 plane. [b] the number of LD-MRS LiDAR frames. [c] the number of individual scans of objects.

with rich information and a considerable measurement range in clear weather conditions. To handle this, some datasets have radar [9, 35, 10] installed, despite the much lower resolution than LiDAR. The proposed dataset publishes the ambient (near infrared) and reflectivity images besides the LiDAR point clouds (see Table 1), which has extreme low-light sensitivity and are robust within poor illumination conditions and adverse weather.

Data diversity within any dataset helps the generation of more universal trained models that can operate successfully under a variety of scenarios. Some related work considers the diversity in their dataset curation [44, 30, 35, 10], but fail to collect data under diverse conditions over the same driving route (see Table 1), *e.g.*, traffic level, times of day, weathers, *etc*. The proposed dataset has a wide range of data diversity via collection over the same repeated route under varying conditions.

Ground truth depth is not present in some seminal autonomous driving datasets, *e.g.*, Stanford Track Collection [53], Sydney Urban Objects [52], Cityscapes [12], Oxford RobotCar [44], LiVi-Set [11], nuScenes [10] and H3D [49]. Due to this limitation, they can only be applied for unsupervised and semi-supervised depth estimation methods [28, 67]. In view of this, our proposed dataset contains ground truth depth at a higher resolution than all previous datasets (Table 1), which is applicable for both supervised and semi-supervised depth estimation tasks.

2.2. Monocular Depth Estimation

Monocular depth estimation aims at recovering a dense depth map for each pixel using a single RGB image as input.

Self-supervised methods harness the monocular RGB image sequences [70, 27, 3, 4, 61], stereo pairs [23, 65, 28, 60, 64] or synthetic data [2, 36] for training. Subsequently, multi-frame architectures were introduced [57, 66, 50, 58, 13, 68, 61], which leverages the temporal information at test time, to improve the quality of the predicted depth. The same losses used during training can be applied to test frames to update the weights. However, additional calculations for multiple forward and backward process on a set of test frames are required which incur additional computation.

Other work concentrates on multi-view stereo (MVS), which operates on unordered image sets [45, 40, 42, 33, 37, 14, 63, 62, 61]. Not requiring the ground truth depth and camera poses during training, self-supervised MVS methods [42, 33, 37, 14, 63, 62, 61] leverage cost volumes to process sequences of frames at test time. Compared with the base method of MVS, these methods can predict the depth using images captured by moving cameras and do not need camera poses during training time.

Supervised methods utilise ground truth depth from depth sensors, *e.g.*, LiDAR [38, 32, 21, 4, 18] and RGB-D cameras [17, 16], to improve the supervision feedback during learning. As with many areas of contemporary computer vision, CNN based architectures [17, 16, 56] generally offer state-of-the-art performance. Thereafter, residual-learning-based methods [31, 39, 69] are proposed to learn the transform relation between colour images and their corresponding maps, therefore leveraging deeper architectures than previous works with higher resultant accuracy. However, such methods are limited both by ground truth dataset availability and the fidelity (resolution) of the ground truth depth information provided.

Overall, one of key challenges within contemporary autonomous driving task evaluation is the lack of high fidelity (vertical resolution) depth datasets in order to facilitate effective evaluation of geometric scene understanding tasks, such as monocular depth estimation. Here, based on

the provision of our DurLAR dataset (Section 3), we consider the impact of abundant high-resolution ground truth depth data on three state-of-the-art contemporary monocular depth estimation architectures (MonoDepth2 [27], Depth-hints [60], ManyDepth [61]) through the use of our novel joint supervised/semi-supervised loss formulation (Section 4).

3. The DurLAR Dataset

Compared to existing autonomous driving task datasets (Table 1), DurLAR has the following novel features:

- **High vertical resolution LiDAR** with 128 channels, which is twice that of any existing datasets (Table 1), full 360°depth, range accuracy to ± 2 cm at 20-50m.
- **Ambient illumination (near infrared) and reflectivity panoramic imagery** are made available in the Mono16 format (2048×128 resolution), with this being only dataset to make this provision (Table 1).
- **No rolling shutter effect**, as our flash LiDAR captures all 128 channels simultaneously.
- **Ambient illumination data** is recorded via an onboard lux meter, which is again not available in previous datasets (Table 1).
- **High-fidelity GNSS/INS** available via an onboard OxTS navigation unit operating at 100 Hz and receiving position and timing data from multiple GNSS constellations in addition to GPS.
- **KITTI data format** adopted as the *de facto* dataset format such that it can be parsed using both the DurLAR development kit and existing KITTI-compatible tools.
- **Diversity over repeated locations** such that the dataset has been collected under diverse environmental and weather conditions over the same driving route with additional variations in the time of day relative to environmental conditions (*e.g.* traffic, pedestrian occurrence, ambient illumination, see Table 1).

3.1. Sensor Setup

The dataset is collected using a Renault Twizy vehicle (Figure 2) equipped with the following sensor configuration (as illustrated in Figure 3):

- **LiDAR**: Ouster OS1-128 LiDAR sensor with 128 channels vertical resolution, 865 nm laser wavelength, 100 m @ >90% detection probability and 120 m @ >50% detection probability (100 klx sunlight, 80% Lambertian reflectivity, 2048 @ 10 Hz rotation rate mode), 0.3 cm range resolution, 360°horizontal FOV and 45°(+22.5°to -22.5°) vertical FOV, mounted height ~ 1.62 m.
- **Stereo Camera**: Carnegie Robotics MultiSense S21 stereo camera with grayscale, colour, and IR enhanced imagers, 0.4 m minimum range, 2048 × 1088 @ 2MP resolution, up to 30 FPS frame rate and 115°× 68°FOV, 21 cm baseline, factory calibrated, mounted height ~ 1.42 m.
- **GNSS/INS**: OxTS RT3000v3 global navigation satellite and inertial navigation system, with 0.03°pitch/roll accuracy, 0.1-1.5 m position accuracy, 0.15°slip angle accuracy, 250 Hz maximum data output rate, supporting positioning from GPS, GLONASS, BeiDou, Galileo, PPP and SBAS constellations.
- **Lux Meter**: Yocto Light V3, a USB ambient light sensor (lux meter), measuring ambient light up to 100,000 lux, hence indirectly representing the conditions of the external environment via ambient illumination conditions.

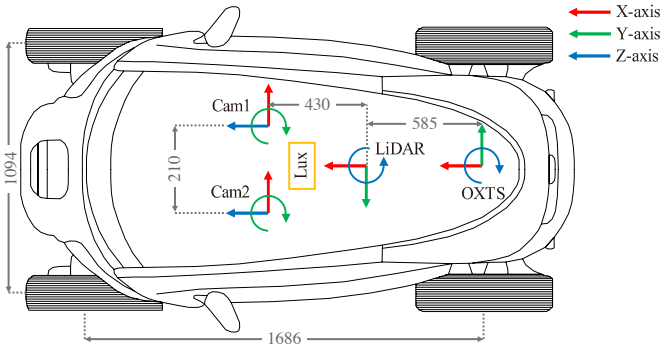

Figure 3: **Sensor placements**, top view. All coordinate axes follow the right-hand rule (sizes in mm).

3.2. Data Collection and Description

To ensure the dataset has diverse weather and varying density of pedestrian and traffic occurrences, we collect the data over a variety of conditions. These includes different types of environments, times of day, weather and repeated locations along the test route with data collected for the key time periods and environments shown in Table 2. As shown

Figure 4: **The route** (blue curves) used for dataset collection showing a variety of driving environments.

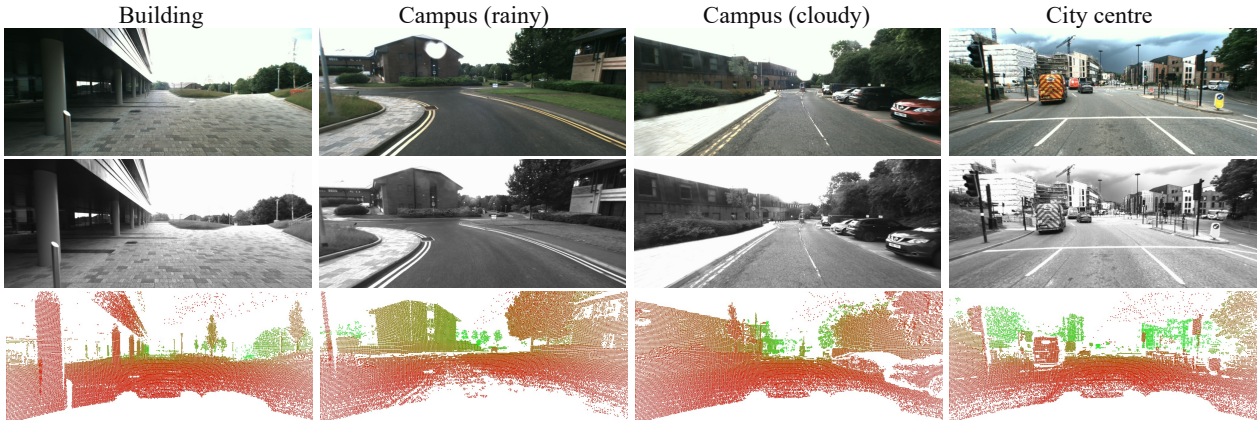

Figure 5: Examples from DurLAR which demonstrate the diversity in our dataset. From top to bottom, RGB left camera images (top), grayscale right camera images (centre) and LiDAR point cloud (bottom). The point cloud is projected onto the 2D image plane using the LiDAR-to-left-camera external calibration, and the colour varies with the distance from the LiDAR (near:=red $\rightarrow$ far:=green).

	Avg. Speed	Day.	Peak times	Night
City	20.4 km/h	[3] | [3]	[3] | [3]	[2] | [3]
Campus	26.4 km/h	[1] | [1]	[1] | [2]	[1] | [1]
Residential	31.2 km/h	[1] | [2]	[2] | [2]	[1] | [1]
Suburb	43.6 km/h	[1] | [1]	[1] | [1]	[1] | [1]

Table 2: Key time periods and environmental conditions. The value is expressed in the form of [traffic density] | [population density], using a qualitative scale of [3 - high, 2 - normal, 1 - low].

in Figure 4 and Figure 5, our dataset mainly contains suburban, highway, city centre and campus areas.

All the data is provided in the *de facto* KITTI data formats, with the exception of the ambient light data (lux) which is not provided by KITTI and is hence published in a simple plain text format with aligned timestamp.

3.3. Ambient and Reflectivity Panoramic Imagery

The proposed DurLAR dataset is the first autonomous driving task dataset to additionally provide high-resolution ambient and reflectivity panoramic 360-degree imagery. The ambient imagery can be captured even in low light conditions (near infrared, 800-2500 nm), while the reflectivity imagery pertains to the material property of the scene object and its reflectivity of the 850 nm LiDAR signal in use (Ouster OS1-128). These characteristics, combined with a superior vertical resolution when compared to other datasets, enable these images to offer great benefit when dealing with unfavourable illumination conditions and coherent scene object identification.

Ambient images offer day/night scene visibility in the near-infrared spectrum. The photon counting ASIC (Application Specific Integrated Circuit) of our sensor has particularly strong illumination sensitivity, so that the ambient images can be captured even in low light conditions. This is extremely practical in designing techniques that are specifically appropriate for adverse illumination conditions, such as nocturnal and adverse weather conditions.

Reflectivity images contain information indicative of the material properties of the object itself and offer good consistency across illumination conditions and range. However, the Ouster OS1-128 LiDAR does not collect the true reflectivity data directly due to sensor limitations. Instead, an estimation of the reflectivity data is used to calculate the reflectivity images from the LiDAR intensity and range data. LiDAR intensity is the return signal strength of the laser pulse that recorded the range reading. According to the inverse square law (Equation (1)) for Lambertian objects in the far field, the intensity per unit area varies inversely proportional to the square of the distance [48],

$$I = \frac{S}{4\pi r^2}, \tag{1}$$

where I is the intensity, r is the range (namely the distance of the object to the sensor) and S is the source strength.

The calculation of reflectivity assumes that it is proportional to the source strength, which is also proportional to the product of intensity and the square of the range,

$$\text{Reflectivity} \propto S \propto Ir^2. \tag{2}$$

Exemplar ambient (near infrared) and reflectivity panoramic imagery is shown in Figure 6. In Figure 6 (a) and (c), clouds and shadows of objects can be distinguished (expressed as shades of grayscale). These pictures are very close to the images of grayscale or RGB camera. In Figure 6 (b) and (d), the reflectivity of the same object or material will remain constant regardless of the distance to the sensor, weather, light illumination and other conditions, since reflectivity is the intrinsic property of the object itself. The

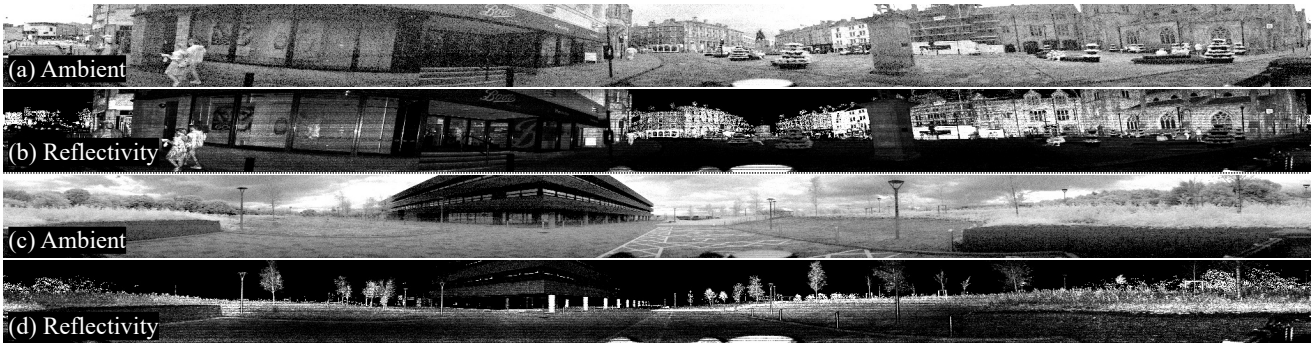

Figure 6: Example of ambient (near infrared) and reflectivity panoramic images.

pillars of the building (Figure 6 (d)) have almost the same reflectivity (*i.e.* the same white colour in the figure) regardless of their distance to the LiDAR sensor.

3.4. Calibration and Synchronisation

LiDAR-to-camera calibration is performed using [15, 6]. With the custom calibration pattern shown in Figure 7, the calibration procedure is composed of two stages. Firstly, a pair of two ArUco markers are detected from the left frame of the stereo camera such that the transformation matrix $[R|t]$, containing rotation R and translation t parameters, between the camera and the centre of the ArUco marker can be calculated (as shown in the overlays of Figure 8). Secondly, the edges of the orientated calibration boards are identified in the corresponding LiDAR data frame projection by orientated edge detection. Finally, the optimal rigid transformation between the LiDAR and the camera is found using RANSAC based optimisation [15].

Stereo camera calibration is based on the manufacturer factory instructions for intrinsic and extrinsic settings. Calibration of the GNSS/INS is performed using the manufacturers recommended approach. The GNSS/INS with respect to the LiDAR is registered following [19].

All sensor synchronisation is performed at a rate of 10 Hz, using Robot Operating System (ROS, version Noetic) timestamps operating over a Gigabit Ethernet backbone to a common host (Intel Core i5-6300U, 16 GB RAM).

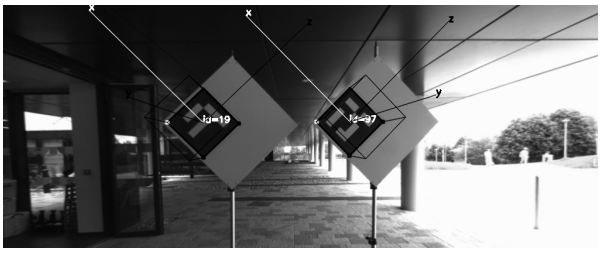

Figure 7: Camera to LiDAR custom calibration pattern with extrinsic parameter estimation overlay shown.

Figure 8: Illustrative LiDAR 3D point cloud overlay onto the right stereo image (colour) using the calibration obtained.

4. Monocular Depth Estimation

Leveraging the higher vertical LiDAR resolution of our DurLAR dataset, we adopt monocular depth estimation as an illustrative benchmark task.

We select ManyDepth [61] as a leading approach for monocular depth estimation as it offers state-of-the-art performance on the leading KITTI [24] and Cityscapes [12] benchmarks. Whilst ManyDepth [61] is a self-supervised approach, here we seek to leverage the availability of high-fidelity depth within DurLAR via the introduction of a secondary supervised loss term to formulate a novel supervised/self-supervised loss formulation. As a result, we can assess the impact of the availability of abundant ground truth depth at training time on the performance of this leading contemporary approach.

To these ends, we introduce the reverse Huber (Berhu) loss $\mathscr{L}_{\text{Berhu}}$ [71] as our supervised depth loss term, due to its effectiveness in smoothing and blurring depth prediction edges on object boundaries:

$$\mathscr{L}_{\text{Berhu}}\ (d, d^*) = \begin{cases} |d - d^*| & \text{if } |d - d^*| \leq \delta, \\ \frac{(d-d^*)^2+\delta^2}{2\delta} & \text{if } |d - d^*| > \delta, \end{cases} \quad (3)$$

where d is the predicted depth, d^* is the ground truth depth, and δ stands for the threshold. If $|d - d^*| \leq \delta$, the Berhu loss is equal to $\mathscr{L}_1$; else, it acts approximately as $\mathscr{L}_2$.

We hence construct a joint supervised/semi-supervised

Dataset	Method	+S	W × H	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
KITTI [24]	ManyDepth (MR) [61]	×	640 × 192	0.098	0.770	4.459	0.176	0.900	0.965	**0.983**
	ManyDepth (HR) [61]	×	1024 × 320	**0.093**	**0.715**	**4.245**	**0.172**	**0.909**	**0.966**	**0.983**
Cityscapes [12]	ManyDepth [61]	×	416 × 128	0.114	1.193	6.223	0.170	0.875	0.967	0.989
DurLAR	Depth-hints [60]	×	640 × 192	0.122	1.070	4.148	0.211	0.870	0.946	0.972
	Depth-hints [60]	✓	640 × 192	0.121	1.109	4.121	0.210	0.874	0.946	0.972
	MonoDepth2 [27]	×	640 × 192	0.111	1.114	4.002	0.187	0.895	0.960	0.981
	MonoDepth2 [27]	✓	640 × 192	0.108	1.010	3.804	0.185	0.898	0.963	0.982
	ManyDepth (MR) [61]	×	640 × 192	0.115	1.227	4.116	0.186	0.892	0.962	0.982
	ManyDepth (MR) [61]	✓	640 × 192	0.109	**0.936**	3.711	0.176	0.895	0.964	0.984
	ManyDepth (HR) [61]	×	1024 × 320	0.109	1.111	3.875	0.177	0.901	0.966	0.984
	ManyDepth (HR) [61]	✓	1024 × 320	**0.104**	**0.936**	**3.639**	**0.171**	**0.906**	**0.969**	**0.986**

Table 3: Performance comparison over the KITTI Eigen split [26], Cityscapes [12] (self-supervised only) and DurLAR datasets (+S, joint supervised/self-supervised (✓) *v.s.* self-supervised (×)). MR and HR stand for medium and high resolution of training models (as originally defined in [61]). Depth evaluation metrics are shown in the top row. Red refers to superior performances indicated by low values, and green refers to superior performance indicated by a higher value. The best results in KITTI and DurLAR are in **bold**; the second best in DurLAR are underlined.

version of ManyDepth [61], adding $\mathscr{L}_{\text{Berhu}}$ into the original ManyDepth loss function, as shown in Equation (4):

$$\mathscr{L} = (1 - M)\mathscr{L}_p + \mathscr{L}_{\text{consistency}} + \mathscr{L}_{\text{smooth}} + \mathscr{L}_{\text{Berhu}}\,, \quad (4)$$

where $\mathscr{L}_p$ is the photometric reprojection error and $\mathscr{L}_{\text{smooth}}$ is the smoothness loss, from [27, 61]. $\mathscr{L}_{\text{consistency}}$ is the consistency loss, as implemented from [61].

For an extended comparison, we similarly introduce this additional supervised depth loss via this additional Berhu loss term to the contemporary MonoDepth2 [27] and Depth-hints [60] approaches leaving the remainder of the architectures unchanged.

We specify a randomly generated data split for the DurLAR dataset as well, comprising 90k training frames, 5k validation frames and 5k test frames for our evaluation.

5. Evaluation Results

Training was performed with all learning parameters set as per the original works [27, 61, 60], with Berhu threshold $\delta = 0.2$, on a Nvidia Tesla V100 GPU over 20 epochs.

5.1. Quantitative Evaluation

The varying performance of self-supervised depth estimation between the KITTI [24], Cityscapes [12] and proposed DurLAR dataset illustrates the varying levels of challenge and complexity afforded by variations within the datasets (Table 3, records with × in the +S column)

However, within our evaluation on the DurLAR dataset, we consistently observe superior performance (lower RMSE, higher accuracy, *etc,* Table 3) with the use of additional depth supervision (*i.e.* joint supervised/semi-supervised loss, see Table 3 - records with ✓in the +S column) across all three monocular depth estimation approaches considered and show overall state-of-the-art performance on monocular depth estimation using our joint supervised/self-supervised ManyDepth variant (DurLAR, Table 3 - as highlighted in bold).

5.2. Qualitative Evaluation

To qualitatively illustrate the difference between self-supervised and joint supervised/self-supervised ManyDepth with the addition of depth loss, we show exemplars highlighting areas of superior depth estimation (Figure 9).

Within these examples, we can see a clearer contour edge of the bus and resolution of the upper LED display board on the vehicle (Figure 9, top - self-supervised *v.s.* supervised/self-supervised). Furthermore, we see improved depth resolution of the building (Figure 9, middle - self-supervised *v.s.* supervised/self-supervised) whereby additional depth supervision enables the technique to correctly estimate the depth of the supporting building pillars and is even able to resolve the depth of the short stainless steel stub in the foreground. Finally, we can see improved estimation and clarity of both vehicle and pedestrian depth within a crowded urban scene (Figure 9, bottom - self-supervised *v.s.* supervised/self-supervised).

Furthermore, we conduct additional comparative cross-training experiments to explore training on DurLAR, KITTI or KITTI/DurLAR combined whilst evaluating on a novel KITTI/DurLAR union split (Table 4). Our KITTI/DurLAR union training/testing data split presents a challenging evaluation task that is more diverse, with 694 test frames each from KITTI and DurLAR, to measure the overall performance across both datasets.

5.3. Ablation Study

Our ablation study shows the side-by-side impact of our joint supervised/unsupervised loss formulation in addition to the performance impact of high-fidelity depth (higher vertical LiDAR resolution).

Supervised depth: We train the ManyDepth [61] with

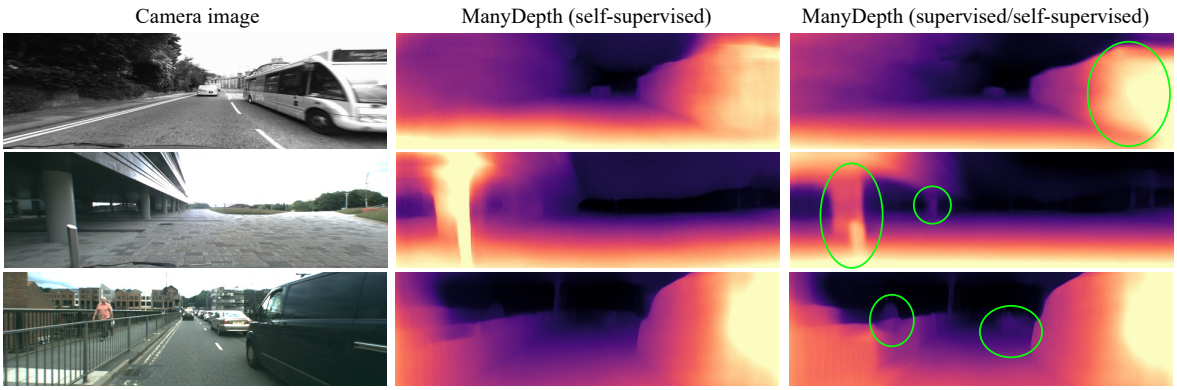

Figure 9: Comparison of monocular depth estimation results with areas of improvement highlighted with the use of depth supervision (green).

and without the Berhu loss (Equation 3), such that we can compare the original self-supervised performance with that of additional depth supervision (Table 5, 128/-S *v.s.* 128/+S).

Ground truth depth resolution: We simulate a reduction in vertical ground truth depth resolution by subsampling the depth values present by 50% (64 channels) and 75% (32 channels) along the vertical axis of the LiDAR ground truth projection. From Table 5, we can see superior performance from our joint supervised/unsupervised loss formulation (128/-S *v.s.* 128/+S) and from higher vertical resolution LiDAR (32/64 *v.s.* 128/-S).

Train	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
K	0.159	1.536	5.101	0.244	0.798	0.923	0.963
D	0.189	1.764	5.580	0.264	0.758	0.908	0.959
K+D	0.188	1.941	5.182	0.262	0.769	0.912	0.958
D+K	**0.151**	**1.123**	**4.744**	**0.233**	**0.805**	**0.927**	**0.967**

Table 4: **Cross-dataset tests** of ManyDepth [51] with the training configuration (K) KITTI only, (D) DurLAR only, (K+D) KITTI then fine-tuning with DurLAR, (D+K) DurLAR then fine-tuning with KITTI. δ_1, δ_2 and δ_3 refers to $\delta < 1.25$, $\delta < 1.2^2$ and $\delta < 1.25^3$ respectively.

vRes	Abs Rel	Sq Rel	RMSE	RMSE log	δ_1	δ_2	δ_3
32/+S	0.115	**0.908**	3.677	0.179	0.888	0.966	0.985
64/+S	0.107	0.918	3.735	0.175	0.895	0.967	**0.986**
128/-S	0.109	1.111	3.875	0.177	0.901	0.966	0.984
128/+S	**0.104**	0.936	**3.639**	**0.171**	**0.906**	**0.969**	**0.986**

Table 5: **Ablation results on ManyDepth** [61]. vRes := the vertical resolution of LiDAR ground truth depth. ±S := supervised/self-supervised (+S) and self-supervised ManyDepth (-S) for consistency with Table 3.

6. Conclusion

In this paper, we present a high-fidelity 128-channel 3D LiDAR dataset with panoramic ambient (near infrared) and reflectivity imagery for autonomous driving applications (DurLAR). In addition, we present the exemplar benchmark task of depth estimation task whereby we show the impact of higher resolution LiDAR as a means to the supervised extension of leading contemporary monocular depth estimation approaches [27, 60, 61].

DurLAR, is a novel large-scale dataset comprising contemporary high-fidelity LiDAR, stereo/ambient/reflectivity imagery, GNSS/INS and environmental illumination information under repeated route, variable environment conditions (in the *de facto* KITTI dataset format). It is the first autonomous driving task dataset to additionally comprise usable ambience and reflectivity LiDAR obtained imagery (2048×128 resolution).

In our sample monocular depth estimation task, we show superior performance can be achieved by leveraging the high resolution LiDAR resolution afforded by DurLAR via the secondary introduction of an additional supervised loss term for depth. This is demonstrated across three state-of-the-art monocular depth estimation approaches [27, 60, 61]. We show that the recent availability of abundant high-resolution ground truth depth from sensors such as those used in DurLAR enable new research possibilities for supervised learning within this domain.

Further work will consider the provision of additional dataset annotation spanning object, semantic and geometric scene information. Future application utilising the ambient and reflectivity imagery will be explored.

Acknowledgements: This work made use of the facilities of the N8 Centre of Excellence in Computationally Intensive Research (N8 CIR) provided and funded by the N8 research partnership and EPSRC (Grant No. EP/T022167/1). The Centre is co-ordinated by the Universities of Durham, Manchester and York.

References

[1] I. Alhashim and P. Wonka. High quality monocular depth estimation via transfer learning. *arXiv preprint arXiv:1812.11941*, 2018. 1

[2] A. Atapour-Abarghouei and T. Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2800–2810. IEEE, June 2018. 3

[3] A. Atapour-Abarghouei and T. Breckon. Monocular segment-wise depth: Monocular depth estimation based on a semantic segmentation prior. In *IEEE Int. Conf. Image Processing*, pages 4295–4299. IEEE, September 2019. 3

[4] A. Atapour-Abarghouei and T. Breckon. To complete or to estimate, that is the question: A multi-task depth completion and monocular depth estimation. In *Int. Conf. 3D Vision*, pages 183–193. IEEE, September 2019. 3

[5] 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 *IEEE Int. Conf. Computer Vision*, pages 9296–9306, 2019. 1, 2, 3

[6] J. Beltrán, C. Guindel, and F. García. Automatic Extrinsic Calibration Method for LiDAR and Camera Sensor Setups. *arXiv preprint arXiv:2101.04431*, 2021. 6

[7] S. F. Bhat, I. Alhashim, and P. Wonka. Adabins: Depth estimation using adaptive bins. pages 4009–4018, 2021. 1

[8] J. W. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M. M. Cheng, and I. Reid. Unsupervised Scale-Consistent Depth and Ego-Motion Learning From Monocular Video, 2019. 1

[9] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide. Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 11679–11689, 2020. 1, 3

[10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. Nuscenes: A Multimodal Dataset for Autonomous Driving. *IEEE Conf. Computer Vision and Pattern Recognition*, pages 11618–11628, 2020. 1, 2, 3

[11] Y. Chen, J. Wang, J. Li, C. Lu, Z. Luo, H. Xue, and C. Wang. LiDAR-Video Driving Dataset: Learning Driving Policies Effectively. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 5870–5878, 2018. 1, 2, 3

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 3, 6, 7

[13] A. CS Kumar, S. M. Bhandarkar, and M. Prasad. Depthnet: A recurrent neural network architecture for monocular depth prediction. In *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, pages 283–291, 2018. 3

[14] Y. Dai, Z. Zhu, Z. Rao, and B. Li. Mvs^2: Deep unsupervised multi-view stereo with multi-view symmetry. In *Int. Conf. 3D Vision*, pages 1–8, 2019. 3

[15] A. Dhall, K. Chelani, V. Radhakrishnan, and K. M. Krishna. LiDAR-Camera Calibration Using 3D-3D Point Correspondences. *ArXiv e-prints*, 2017. 6

[16] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *IEEE Int. Conf. Computer Vision*, pages 2650–2658, 2015. 3

[17] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Conf. Neural Information Processing Systems (NeurIPS)*, pages 2366–2374, 2014. 3

[18] A. Eldesokey, M. Felsberg, K. Holmquist, and M. Persson. Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2020. 3

[19] A. W. Fitzgibbon. Robust registration of 2d and 3d point sets. *Image and Vision Computing*, 21(13):1145–1153, 2003. 6

[20] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 1

[21] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 3

[22] D. Garg, Y. Wang, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao. Wasserstein Distances for Stereo Disparity Estimation. In *Conf. Neural Information Processing Systems (NeurIPS)*, 2020. 1

[23] R. Garg, V. K. Bg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Euro. Conf. Computer Vision*, pages 740–756. Springer, 2016. 3

[24] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision Meets Robotics: The KITTI Dataset. *Int. J. Robotics Research*, 32(11):1231–1237, 2013. 1, 2, 3, 6, 7

[25] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2012. 1

[26] A. Geiger, P. Lenz, and R. Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 2, 7

[27] C. Godard, O. M. Aodha, M. Firman, and G. Brostow. Digging Into Self-Supervised Monocular Depth Estimation. *IEEE Int. Conf. Computer Vision*, 2019-October:3827–3837, 2019. 1, 2, 3, 4, 7, 8

[28] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised Monocular Depth Estimation With Left-Right Consistency. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 6602–6611, 2017. 3

[29] T. Gruber, M. Bijelic, F. Heide, W. Ritter, and K. Dietmayer. Pixel-Accurate Depth Evaluation in Realistic Driving Scenarios. In *Int. Conf. 3D Vision*, pages 95–105, 2019. 1

[30] T. Gruber, F. Julca-Aguilar, M. Bijelic, and F. Heide. Gated2Depth: Real-Time Dense LiDAR from Gated Images. In *IEEE Int. Conf. Computer Vision*, pages 1506–1516, 2019. 1, 2, 3

[31] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3

[32] L. He, C. Chen, T. Zhang, H. Zhu, and S. Wan. Wearable depth camera: Monocular depth estimation via sparse optimization under weak supervision. *IEEE Access*, 6:41337–41345, 2018. 3

[33] Y. Hou, J. Kannala, and A. Solin. Multi-view stereo by temporal nonparametric fusion. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2651–2660, 2019. 3

[34] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. https://level-5.global/level5/data/, 2020. 1, 2, 3

[35] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft Level 5 Perception Dataset 2020. https://level5.lyft.com/dataset/, 2019. 1, 2, 3

[36] F. Khan, S. Hussain, S. Basak, J. Lemley, and P. Corcoran. An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data. *Neural Networks*, 142:479–491, 2021. 3

[37] T. Khot, S. Agrawal, S. Tulsiani, C. Mertz, S. Lucey, and M. Hebert. Learning unsupervised multi-view stereopsis via robust photometric consistency. In *IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2019. 3

[38] Y. Kuznietsov, J. Stuckler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 6647–6655, 2017. 3

[39] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Int. Conf. 3D Vision*, pages 239–248, 2016. 3

[40] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang. Learning for disparity estimation through feature constancy. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. 3

[41] Y. Liao, J. Xie, and A. Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint arXiv:2109.13410*, 2021. 2, 3

[42] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, and J. Kautz. Neural rgb→d sensing: Depth and uncertainty from a video camera. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 10986–10995, 2019. 3

[43] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, and P. Newman. Real-Time Kinematic Ground Truth for the Oxford RobotCar Dataset. *arXiv preprint arXiv: 2002.10152*, 2020. 1

[44] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The Oxford RobotCar dataset. *Int. Journal of Robotics Research*, 36(1):3–15, 2017. 1, 3

[45] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 3

[46] M. Menze and A. Geiger. Object Scene Flow for Autonomous Vehicles. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 2

[47] Ouster. Webinar: Digital vs Analog Lidar. https://ouster.com/resources/webinars/digital-vs-analog-lidar/. 2

[48] Ouster. Webinar: How to understand lidar performance: range, precision, and accuracy. https://go.ouster.io/webinar/how-to-understand-lidar-performance-range-precision-accuracy/. 5

[49] A. Patil, S. Malla, H. Gang, and Y. T. Chen. The H3D Dataset for Full-Surround 3D Multi-Object Detection and Tracking in Crowded Urban Scenes. In *IEEE Int. Conf. Robotics and Automation*, volume 2019-May, pages 9552–9557, 2019. 1, 2, 3

[50] V. Patil, W. Van Gansbeke, D. Dai, and L. Van Gool. Don't forget the past: Recurrent depth estimation from monocular video. *Robotics and Automation Letters*, 5(4):6813–6820, 2020. 3

[51] K. Qian, S. Zhu, X. Zhang, and L. E. Li. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary LiDAR and Radar Signals. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 444–453, 2021. 2

[52] A. Quadros, J. Underwood, and B. Douillard. Sydney Urban Objects Dataset. 2013. 1, 2, 3

[53] A. Teichman, J. Levinson, and S. Thrun. Towards 3D Object Recognition via Classification of Arbitrary Object Tracks. In *IEEE Int. Conf. Robotics and Automation*, pages 4034–4041, 2011. 1, 2, 3

[54] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, K. Lau, C. Oakley, M. Palatucci, V. Pratt, P. Stang, S. Strohband, C. Dupont, L. E. Jendrossek, C. Koelen, C. Markey, C. Rummel, J. van Niekerk, E. Jensen, P. Alessandrini, G. Bradski, B. Davies, S. Ettinger, A. Kaehler, A. Nefian, and P. Mahoney. Stanley: The Robot That Won the DARPA Grand Challenge. *J. Field Robotics*, 23(9):661–692, 2006. 2

[55] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger. Sparsity Invariant CNNs. In *Int. Conf. 3D Vision*, pages 11–20, 2017. 1

[56] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 5038–5047, 2017. 3

[57] J. Wang, G. Zhang, Z. Wu, X. Li, and L. Liu. Self-supervised joint learning framework of depth estimation via implicit cues. *arXiv preprint arXiv:2006.09876*, 2020. 3

[58] R. Wang, S. M. Pizer, and J.-M. Frahm. Recurrent neural network for (un-)supervised learning of monocular video visual odometry and depth. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 5555–5564, 2019. 3

[59] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. Van Der Maaten, M. Campbell, and K. Q. Weinberger. Anytime Stereo Im-

age Depth Estimation on Mobile Devices. *IEEE Int. Conf. Robotics and Automation*, pages 5893–5900, 2019. 1

[60] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov. Self-supervised monocular depth hints. In *IEEE Int. Conf. Computer Vision*, October 2019. 2, 3, 4, 7, 8

[61] J. Watson, O. Mac Aodha, V. Prisacariu, G. Brostow, and M. Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2021. 2, 3, 4, 6, 7, 8

[62] F. Wimbauer, N. Yang, L. von Stumberg, N. Zeller, and D. Cremers. Monorec: Semi-supervised dense reconstruction in dynamic environments from a single moving camera. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 6112–6122, 2021. 3

[63] Z. Wu, X. Wu, X. Zhang, S. Wang, and L. Ju. Spatial correspondence with generative adversarial network: Learning depth from monocular videos. In *IEEE Int. Conf. Computer Vision*, pages 7494–7504, 2019. 3

[64] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao. Structure-guided ranking loss for single image depth prediction. In *IEEE Conf. Computer Vision and Pattern Recognition*, 2020. 3

[65] J. Xie, R. Girshick, and A. Farhadi. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Euro. Conf. Computer Vision*, pages 842–857. Springer, 2016. 3

[66] J. Xie, C. Lei, Z. Li, L. E. Li, and Q. Chen. Video depth estimation by fusing flow-to-depth proposals. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pages 10100–10107. IEEE, 2020. 3

[67] Z. Yin and J. Shi. GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1983–1992, 2018. 3

[68] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, and Y. Yan. Exploiting temporal consistency for real-time video depth estimation. In *IEEE Int. Conf. Computer Vision*, pages 1725–1734, 2019. 3

[69] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang. Joint task-recursive learning for semantic segmentation and depth estimation. In *Euro. Conf. Computer Vision*, pages 235–251, 2018. 3

[70] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 1851–1858, 2017. 2, 3

[71] L. Zwald and S. Lambert-Lacroix. The BerHu Penalty and the Grouped Effect. *arXiv preprint arXiv:1207.6868*, 2012. 6