Too Tiny to See: Hazardous Obstacle Detection Dataset and Evaluation

Topi Miekkala[1] Samuel Brucker[2] Stefanie Walz[2] Filippo Ghilotti[2]
Andrea Ramazzina[3] Dominik Scheuble[3] Pasi Pyykönen[1] Mario Bijelic[2,4] Felix Heide[2,4]
[1]VTT Technical Research Centre of Finland [2]Torc Robotics [3]Mercedes-Benz [4]Princeton University
https://light.princeton.edu/2T2S

Abstract

We introduce a novel dataset and evaluation approach for long-range depth prediction of small objects that enables consistent comparison across direct time-of-flight (ToF) sensors and learned depth estimation methods. In autonomous driving, accurate depth perception is essential for identifying and locating surrounding elements and determining safe driving paths. Traditional depth metrics focus on distance accuracy but fail to evaluate a key factor at long ranges: distinguishing small, slightly elevated structures from the ground — crucial for anticipating obstacles and making safe driving decisions. At far distances, image-based systems suffer from resolution limitations that tend to oversmooth the ground plane, causing elevated objects to be mistaken as texture patterns on the surface. Conversely, scanning LiDAR systems may return only a single point from an elevated object due to steep incident angles and sparse returns, preventing accurate differentiation from the ground. This hampers a fair comparison of object presence and shape. To address this, we propose a framework that evaluates how well the estimated point clouds preserve semantic content relative to ground-truth data. We leverage neural network-based feature extraction to assess structural similarity, enabling a modality-agnostic evaluation of object-level fidelity. Our method also supports analysis of the trade-off between resolution and accuracy, investigating performances across sensor types — such as high-resolution cameras versus LiDAR — and conditions, including day and night scenarios. This enables a more comprehensive understanding of the capabilities and limitations of current depth prediction approaches in real-world settings.

1. Introduction

Depth estimation has seen significant progress in recent years, driven by advances in learning-based techniques [30, 37, 51, 79, 81, 86, 89] and increasingly diverse sensing modalities [12, 64, 77]. Approaches leveraging monocular cues [30, 37, 54, 89], stereo image pairs [18, 47, 51, 81, 86], generalized multi views setups [46, 79, 80], depth completion with sparse input [87, 91, 92], cross-spectral fusion [12, 76, 93], and time-of-flight (ToF) sensors [64] have all contributed to increasingly robust depth prediction pipelines. These methods enable dense scene understanding across a wide range of applications including robotics [53, 56], autonomous driving [16, 67, 73], and augmented reality [52, 57, 78]. Despite these strides, the standard evaluation of depth prediction methods remains largely centered on global, pixel-wise error metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Squared Relative Error (Sq Rel), and Absolute Relative Error (Abs Rel) [24, 66, 73]. While effective in measuring overall prediction fidelity, these metrics inherently emphasize regions of high pixel density, typically the dominant planar surfaces in a scene such as roads and walls [19, 40, 41, 68]. Fine grained details, especially small or distant objects, are likely to contribute negligibly to overall error metrics [38], so safety critical elements such as pedestrians, bicycles, or traffic cones [28, 32] can be present in the scene yet exert almost no influence on evaluation outcomes. Because these objects cover only a small image region, they are often oversmoothed in predicted depth maps [55, 60], as sharp depth discontinuities are hard to model and optimization is geared towards standard metrics. However, accurate depth map representation of small objects is critical, as their shape and height provide essential cues for autonomous driving, enabling systems to determine whether an object can be safely driven over or not. To address this gap, we propose a novel evaluation approach that isolates performance on such foreground objects, with a focus on shape preservation. Our method enables a principled comparison between sensing modalities by explicitly considering the trade-off between resolution and distance accuracy. This allows us to assess not only how well a depth map estimates global structure, but more importantly, whether it preserves the geometry of small, distant, and semantically meaningful objects. Furthermore, we build upon prior work in perceptual image [23, 26, 90] and point cloud [2, 25] quality metrics, where learned representations are used to

evaluate structural similarity beyond pixel-wise error. Similar to learned perceptual image patch similarity (LPIPS) in the 2D domain [90], we extend this idea to 3D point clouds, using features extracted from neural networks to quantify semantic and geometric similarity between predicted and ground-truth object representations.

In summary, our contributions are as follows:

- We present the first long-range dataset designed to evaluate the detection of very small or hard-to-find objects, addressing the critical challenge of obstacles in autonomous driving.
- We introduce the first metric enabling cross-modality evaluation of relative improvements and trade-offs, capturing the contrast between LiDAR's low resolution but high precision, camera's high resolution but reduced distance accuracy, and hybrid sensors such as gated cameras.
- We provide an evaluation showing that, particularly at long ranges, high-resolution camera-based predictions enable more reliable detection of distant objects, consistent with the behavior captured by our proposed metric.

2. Related Work

Autonomous systems depend on reliable depth perception for planning and safety-critical decisions. Fair benchmarking becomes fundamentally important for this task, yet common metrics often miss thin structures, and occlusions.

Depth Estimation. Camera depth estimation encompasses three main families. Monocular methods are fundamentally limited by scale ambiguity [24], with recent approaches attempting to alleviate this by learning scale priors from single images [29, 33, 37, 48, 50]. Stereo-image methods resolve this scale ambiguity by triangulation [18] and include both classical and learned methods [6, 18, 49, 88]. Unsupervised counterparts [27, 29, 30, 33, 94] have also been presented to exploit multi-view geometry consistency as supervision signal. Camera-LiDAR fusion finally uses LiDAR to set metric scale and for geometrical cues [20, 36, 58, 70, 71, 82, 91], at the cost of strict cross-sensor calibration/synchronization. LiDAR is a common ground truth but struggles in adverse weather. Its measurement accuracy sets the precision limit, yet depth methods predict maps at much higher spatial resolution.

Depth Measurement with Time Of Flight. ToF sensors recover range by timing active illumination. Correlation (AMCW) ToF use flood illumination and phase shifts for per-pixel depth, but are fragile under strong ambient light [35, 42, 43]. Pulsed ToF (scanning LiDARs) attain high range accuracy but lower spatial resolution, and performance degrades in fog/snow due to backscatter [10, 17, 39, 64]. Gated imaging integrates flood-illuminated returns within microsecond windows to suppress backscatter and provide coarse depth [3, 11, 13, 14]. Beyond analytic/Bayesian reconstruction [1, 44, 45, 62, 85], recent learning methods infer depth from gated bursts and multi-view setups [31, 75, 77]. These systems, however, are tailored to gated sensors and can be resolution- or power-limited under bright ambient conditions. Recently, this limitation was mitigated by coupling NIR gated sensing with high-resolution visible-spectrum RCCB imagery [12].

Benchmarking. Fair comparison between sensor-measured and algorithmically-estimated depth requires metrics to reflect both accuracy and perceptual relevance. Standard ones such as MAE, RMSE, ARD and Scale-invariant Log Error are commonly used [24, 34, 73] to benchmark performances of environmental reconstruction algorithms, but ignore semantics, structure, and are uninformative for small-objects geometry. Beyond pixels, 3D-space measures such as Voxel Intersection-over-Union (IoU) and Chamfer Distance [7, 22] are common. Chamfer Distance, used in benchmarking and training of 3D reconstruction algorithms, measures the dissimilarity between point sets as the sum of nearest neighbor distances. Although improvements exist [72], the nearest neighbor formulation remains insensitive to non-uniform point densities and under penalizes missing regions that are semantically important. Voxel occupancy–based scores provide an intuitive measure of volumetric consistency, but are dependent on voxel size and suffer from discretization artifacts, with small structures or fine details being lost at coarse resolutions and noise becoming dominant at finer resolutions.

Feature-Based Metrics. Departing from raw-signal-differences metrics, deep feature–based approaches leverage intermediate CNN activations (e.g., VGG) to assess perceptual similarity, super-resolution, and style transfer [23, 26, 90]. A canonical example is the Learned Perceptual Image Patch Similarity (LPIPS) metric [90], which compares images via distances in deep feature space, capturing semantics and structure better than raw-signal differences. Such deep feature–based measures have also been applied to image retrieval [4, 5], face verification [63], and cross-modal tracking and localization. In 3D perception, related concepts appear in large scale place recognition with point cloud encoders [15, 74], SSIM-inspired quality assessment comparing local geometric and color statistics [2], and no-reference quality prediction on orbit videos with 3D CNNs [25]. However, despite demonstrating that deep feature embeddings can encode meaningful semantic and geometric cues, feature-based similarity metrics remain largely under-explored for depth estimation and 3D hazard detection.

3. Dataset

To study the geometric similarity of small ground-level obstacles at long distances, we present a dataset centered on lost cargo—objects that may appear in a vehicle's path,

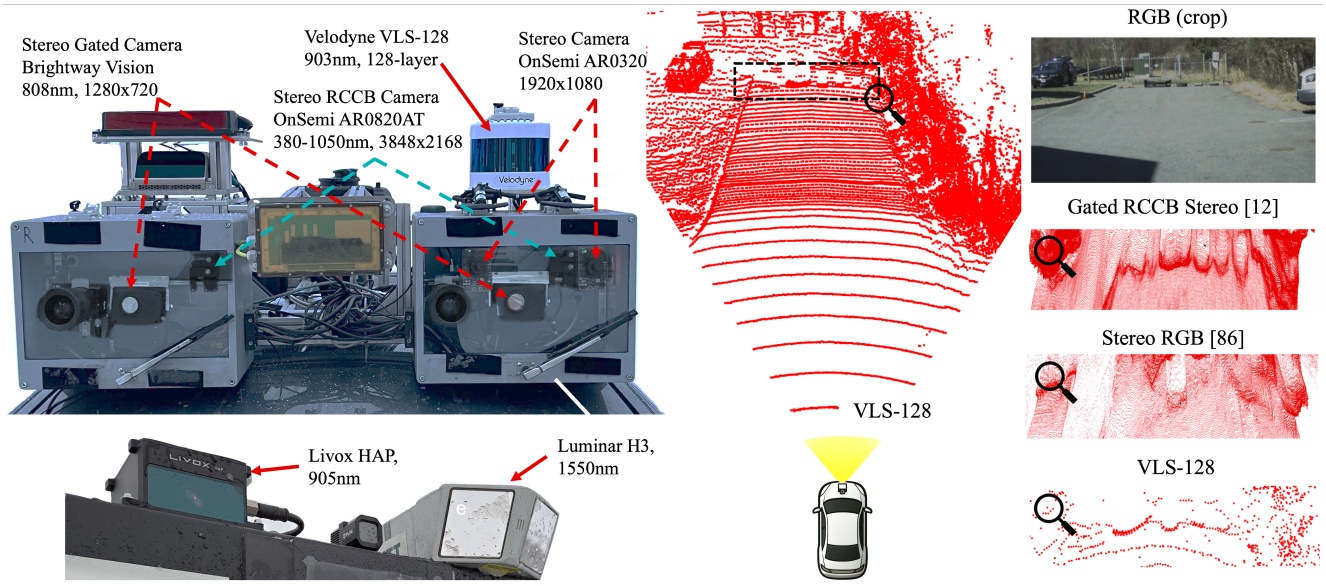

Figure 1. **Sensor Setup and Depth Modalities Limitations.** We present the sensor setup used to capture our dataset, on the left, with the two car-mounted rigs. The center shows a recorded scene with the VLS-128 LiDAR, with lost-cargo objects placed at 30 meters. On the right, the reference RGB image is paired with cropped 3D point clouds from three methods: Gated RCCB Stereo [12] yields dense, well-shaped geometry where the cargo is clearly recognizable; RGB stereo produces over-smoothed clouds in which objects are barely visible; the VLS-128 LiDAR remains very sparse, obscuring small objects and height cues. These perceptual differences, obvious to humans, are poorly captured by standard metrics, motivating our deep similarity metric, which better reflects pointcloud quality.

Sensor	Make	Type	Resolution	FOV (H×V)	Wavelength
RGB Camera	OnSemi	AR0230	1920 × 1024	39.6° × 21.7°	380 – 740 nm
RCCB Camera	OnSemi	AR0820AT	3848 × 2174	52.8° × 28.9°	380 – 1050 nm
Gated Camera	BrightwayVision	BrightEye	1280 × 720	31.1° × 17.8°	808 nm
LiDAR	Velodyne	VLS-128	2000 × 128	360° × 40°	905 nm
LiDAR	Luminar	H3	1660 × 64	120° × 30°	1550 nm
LiDAR	Livox	HAP	—	120° × 25°	905 nm

Table 1. **Sensors Specifications.** We present the list of sensors used in the dataset and their specifications.

such as small items fallen from preceding vehicles and other potential hazards. The dataset consists of scenarios with various lost cargo obstacles manually placed on pavement, and the measurement sensors capturing data at varying distances. The selected lost cargo objects vary in size, material, and shape to capture a range of detection challenges. They include common items such as wooden pallets, car tires, exhaust pipes, and bumpers and motorcyclist dummies ('biker'), with example layouts shown in Figure 1.

Data Collection. We collected a multi-modal dataset to evaluate lost cargo detection performance across diverse sensing modalities using three stereo camera systems and three different LiDAR sensors. The stereo RGB camera features a narrow 0.23 m baseline, operates at 30 Hz, and delivers 1920 × 1024 pixel images with a 39.6° × 21.7° FoV. The RCCB stereo system has a wider 0.75 m baseline and operates at 15 Hz with 3848 × 2176 resolution. Unlike conventional RGB sensors with an RGGB Bayer pattern, RCCB cameras replace the green channels with clear channels, increasing light throughput and improving low-light sensitivity. The gated stereo system is an active sensor comprising two cameras, with a baseline of 0.75m, and a flash laser illuminator. It emits short laser pulses and captures the returning light after predefined delays, producing overlapping range–intensity slices that implicitly encode depth, see [31]. The system has a 31.1° × 17.8° field of view and operates at 120Hz, enabling several slices per frame. LiDAR pointclouds are acquired from the Velodyne VLS-128, a 905nm rotating sensor with 128 scan lines, 40° vertical FoV, 0.11° resolution, and 10 Hz rotation for dense short- to mid-range coverage; the Luminar H3, a 1550 nm fiber-scanning system with >250 m range, operating with an angular resolution of 1660 × 64, and an FoV of 120° × 30°, optimized for detecting small or low-reflectivity objects at long distances; and the Livox HAP, a solid-state LiDAR with a non-repetitive scan pattern, 120° × 25° FoV, up to 150 m range, and high frame rate, enabling dense short- to medium-range coverage with improved point density over time. Table 1 provides detailed overview of technical the specifications. The sensors captured a total of 19 target objects, with 7 recorded in summer conditions and 12 in winter. In summer, 5 distinct objects were used and placed in different orientations and scenes for variety, while in winter 6 distinct objects were used. For each layout, short static sequences were recorded with the vehicle stationary: in winter at four distances (25 m, 50 m, 75 m, 100 m) and in summer at six distances (15 m, 30 m, 45 m, 60 m, 75 m, 90 m). Figure 2 visualizes the data capture procedure

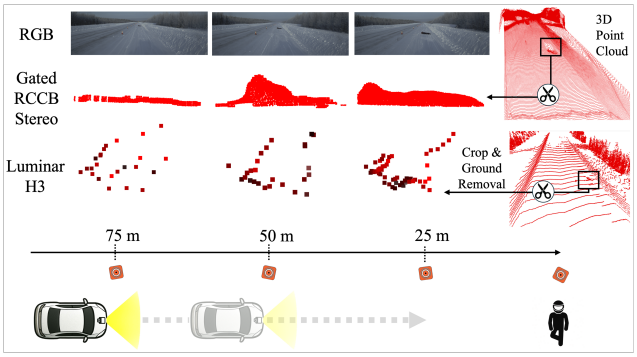

Figure 2. **Safety Critical Obstacle Detections.** 'Biker' target object at distances of 25, 50 and 75 meters in an RGB image, rendered Gated RCCB Stereo [12], and Luminar H3 point cloud.

for an exemplary scene. Further details on sensor configurations, capture modes, and exact recording distances are provided in the Supplementary Material. Inter-sensor calibration was performed to ensure measurement synchronization. The stereo camera systems and the Velodyne VLS-128 were mounted on the same vehicle rig and externally calibrated. The Luminar and Livox LiDARs were mounted on a separate vehicle rig. Synchronization between the two systems is described in the following section.

Ground Truth Creation. For comparison of sensor perception properties at different distances, we created a reference dataset of 3D object models. Each object was either scanned at close range ($\leq$ 1 m) using the commercial photogrammetry app Polycam on an iPhone 12 Pro Max, which combines LiDAR depth sensing with multi-view RGB imagery to generate dense meshes from over 2000 sampled LiDAR and camera scans with example results presented in Figure 3. These meshes serve as high-fidelity ground truth representations. Alternatively, geometry is estimated from a held-out stationary sequence captured with a Livox HAP LiDAR, where 30 point clouds per scene enable dense sampling due to the sensor's non-repeating scan pattern. Here, we acquire a set of N accumulated 3D points for a single object, which are denoted by $\mathcal{P} = \{p_i\}_{i=1}^{N}$, where $p_i = (x_i, y_i, z_i) \in \mathbb{R}^3$ are point coordinates in the LiDAR reference frame. Outliers are removed using a standard deviation filter [61]. For each $p_i \in \mathcal{P}$, let $\mathcal{N}_k(i)$ denote its k nearest neighbors and define the mean neighbor distance

$$d_i \;=\; \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \|p_i - p_j\|_2 \,. \tag{1}$$

We then compute the global mean and variance over $\{d_i\}_{i=1}^{N}$:

$$\mu_d \;=\; \frac{1}{N}\sum_{i=1}^{N} d_i, \qquad \sigma_d^2 \;=\; \frac{1}{N}\sum_{i=1}^{N} (d_i - \mu_d)^2 \,. \tag{2}$$

Points with unusually large d_i are treated as sparse outliers; equivalently, we retain

$$\mathcal{P}' \;=\; \{\, p_i \in \mathcal{P} \;:\; d_i \leq \mu_d + \tau\, \sigma_d \,\}\,, \tag{3}$$

where k is the number of neighbors and $\tau > 0$ is the threshold multiplier.

From the filtered set $\mathcal{P}'$, we reconstruct an initial mesh $\mathcal{M}_{BPA}$ using the Ball Pivoting Algorithm (BPA) [8] with radius r:

$$\mathcal{M}_{BPA} \;=\; \mathrm{BPA}(\mathcal{P}',\, r). \tag{4}$$

Uniform sampling of S points from $\mathcal{M}_{BPA}$ produces the set $\mathcal{Q}$:

$$\mathcal{Q} = \mathrm{Sample}(\mathcal{M}_{BPA}, S). \tag{5}$$

The points $\mathcal{Q}$ are voxelized into an occupancy grid $\mathcal{G} \in \{0,1\}^{X\times Y\times Z}$ with voxel size $v > 0$:

$$\mathcal{G}(x,y,z) = \begin{cases} 1, & \text{if } \exists\, q \in \mathcal{Q} \text{ s.t. } \lfloor q/v \rfloor = (x,y,z), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

A mesh $\mathcal{M}_1 = (V^0, F)$ is then generated from $\mathcal{G}$ using a voxel-to-surface operator Φ_{vox}, where V^0 is the set of vertex positions and F is the set of faces:

$$\mathcal{M}_1 = (V^0, F) = \Phi_{\text{vox}}(\mathcal{G}). \tag{7}$$

The vertex positions $V = \{v_i\}$ are refined by minimizing

$$L_{\text{total}} = \lambda_{\text{c}} L_{\text{chamfer}} + \lambda_{\text{e}} L_{\text{edge}} + \lambda_{\text{l}} L_{\text{laplacian}} + \lambda_{\text{n}} L_{\text{normal}}, \tag{8}$$

where $\lambda_{\text{c}}, \lambda_{\text{e}}, \lambda_{\text{l}}, \lambda_{\text{n}} > 0$ are weighting coefficients.

The Chamfer distance between the current surface samples $\widehat{\mathcal{Q}} = \mathrm{Sample}(\mathcal{M}_1, K)$ and the target point set $\mathcal{P}'$ is

$$L_{\text{chamfer}} = \frac{1}{|\widehat{\mathcal{Q}}|} \sum_{x \in \widehat{\mathcal{Q}}} \min_{y \in \mathcal{P}'} \|x-y\|_2^2 + \frac{1}{|\mathcal{P}'|} \sum_{y \in \mathcal{P}'} \min_{x \in \widehat{\mathcal{Q}}} \|x-y\|_2^2. \tag{9}$$

The edge loss penalizes deviations in edge lengths from those in the initial mesh:

$$L_{\text{edge}} = \sum_{(i,j)\in E} \left(\|v_i - v_j\|_2 - \|v_i^0 - v_j^0\|_2 \right)^2, \tag{10}$$

where E is the set of edges and v_i^0 the initial vertex positions.

The Laplacian loss encourages smoothness by minimizing

$$L_{\text{laplacian}} = \sum_i \left\| v_i - \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} v_j \right\|_2^2, \tag{11}$$

where $\mathcal{N}(i)$ is the one-ring neighborhood of v_i.

The normal consistency loss promotes alignment of normals across adjacent faces,

$$L_{\text{normal}} = \sum_{(f,f')\in \mathcal{A}} \left(1 - \langle n_f, n_{f'} \rangle\right), \tag{12}$$

where $\mathcal{A}$ is the set of adjacent face pairs and n_f the unit normal of f.

The meshes in Fig. 3 were created with the Polycam approach; examples of the Livox-based generation are provided in the Supplementary Material. We use these meshes as per-object ground truth: for each lost cargo layout, one ground truth mesh per object is created and used as the reference across all measurement distances for that layout.

4. Too Tiny To See

This section details the computation steps of our **Too T**iny **To S**ee metric (**2T2S**). First an alignment of point clouds and ground truth meshes and the subsequent processing by higher order neural networks to extract semantic rich embeddings for quantitative comparison.

Initial Point Cloud Alignment. Ground truth objects are represented as $3D$ meshes $\mathcal{M}$ consisting of vertices $\{v_1^o, \ldots, v_{N_v}^o\} \subset \mathbb{R}^3$ and faces $\{f_1, \ldots, f_{N_f}\} \subset \mathbb{N}^3$, where each face indexes three vertices. These meshes are positioned in the nearest measurement point cloud to serve as reference geometry for subsequent comparisons. More details on the automated object placement are provided in the Supplementary Material.

After placing the ground truth objects to the nearest point cloud, we use this pose information to align them to a full point cloud measurement $\mathcal{P}_d$, the ICP algorithm [9] is used to find an alignment to the nearest point cloud image of the measurement sequence. If ICP is computed in a reference sensor frame, we compose the result with the known extrinsic calibration $\mathcal{T}_\mathrm{e}$ from the reference to the evaluated sensor; otherwise we set $\mathcal{T}_\mathrm{e} = \mathcal{I}$, where $\mathcal{I}$ is the identity matrix. The aligned mesh at distance d is then

$$\mathcal{M}_\mathrm{d} = \mathcal{T}_\mathrm{e}\, \mathcal{T}_\mathrm{d}\, \mathcal{M}_0. \tag{13}$$

where $\mathcal{T}_\mathrm{e}$ and $\mathcal{T}_\mathrm{d}$ are homogeneous transformations:

$$\mathcal{T}_\mathrm{d} = \begin{bmatrix} \mathcal{R}_\mathrm{d} & t_\mathrm{d} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \qquad \mathcal{T}_\mathrm{e} = \begin{bmatrix} \mathcal{R}_\mathrm{e} & t_\mathrm{e} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{14}$$

Fine Pose Optimization aims to estimate a small rigid correction between the measured pointcloud P_d and the ground truth mesh with vertices $\{v_j\}$, by minimizing a weighted point-to-mesh distance

$$\mathcal{L}_\mathrm{p2m} = \sum_{i=1}^{N} s_i, \qquad s_i = \left(d_i\, \widehat{\rho}_i + \widetilde{w}_i^x\right) w_i^z \tag{15}$$

with $d_{ij} = \|p_i - v_j\|_2$ being the nearest-vertex distance to the mesh vertices $\{v_j\}$. The density factor $\widehat{\rho}_i$ up-weights well-sampled neighborhoods $\rho_i = \mathrm{dens}(p_i; \lambda)$ in a ball of radius λ and is min-max normalized so that dense, reliable regions get more weight in the pose fit:

$$\widehat{\rho}_i = \frac{\rho_i - \rho_\mathrm{min}}{\rho_\mathrm{max} - \rho_\mathrm{min}} \tag{16}$$

The height weight w_i^z is defined

$$w_i^z = \frac{z_i - z_\mathrm{min}}{z_\mathrm{max} - z_\mathrm{min}}, \tag{17}$$

with z_i being the point height, reduces the influence of near-ground clutter with normalization. Finally, to down-weight boundary artifacts, we define a lateral x-axis vertex weight

$$w_j^x = 1 - \frac{v_j^x - v_\mathrm{min}^x}{v_\mathrm{max}^x - v_\mathrm{min}^x}, \tag{18}$$

assigned to each point through its nearest mesh vertex

$$j^*(i) = \arg\min_j d_{ij}, \qquad d_{ij} = \|p_i - v_j\|_2, \tag{19}$$

and sharpened with weight

$$\widetilde{w}_i^x = \left(w_{j^*(i)}^x\right)^4. \tag{20}$$

Through stochastic gradient descent, we optimize rotation $\mathcal{R}_\mathrm{o}$ and translation t_o of

$$\mathcal{T}_\mathrm{o} = \begin{bmatrix} \mathcal{R}_\mathrm{o} & t_\mathrm{o} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{21}$$

yielding the evaluation mesh

$$\mathcal{M}_\mathrm{e} = \mathcal{T}_\mathrm{o}\, \mathcal{T}_\mathrm{e}\, \mathcal{T}_\mathrm{d}\, \mathcal{M}_0, \tag{22}$$

achieving optimal alignment for quality evaluation. We crop points within distance d_e and apply a z-range filter, yielding the evaluation point cloud $\mathcal{P}_e$.

2T2S Metric. To compare object similarity beyond purely geometric metrics, we employ a deep feature representation learned from $3D$ point clouds. A neural network is trained on a large-scale object classification dataset to capture shape-discriminative features that are invariant to minor geometric distortions and partial observations.

For each evaluated point cloud $\mathcal{P}_\mathrm{e}$, the network produces intermediate feature matrices

$$Y_\mathrm{e}^{(l)} \in \mathbb{R}^{N \times F_l}, \quad l = 1, \ldots, L, \tag{23}$$

from L latent layers, where N is the number of latent samples and F_l is the dimensionality of the l-th feature space. The sampled ground truth mesh $\mathcal{M}_0$ is uniformly sampled to a point-set $\mathcal{P}_0$, for input compatibility with the feature encoder, and passed through the same network to obtain the corresponding matrices

$$Y_0^{(l)} \in \mathbb{R}^{N \times F_l}. \tag{24}$$

Each feature matrix is first normalized to unit length on a per-sample basis. For each feature dimension $f = 1, \ldots, F_l$, we then compute the the one-dimensional

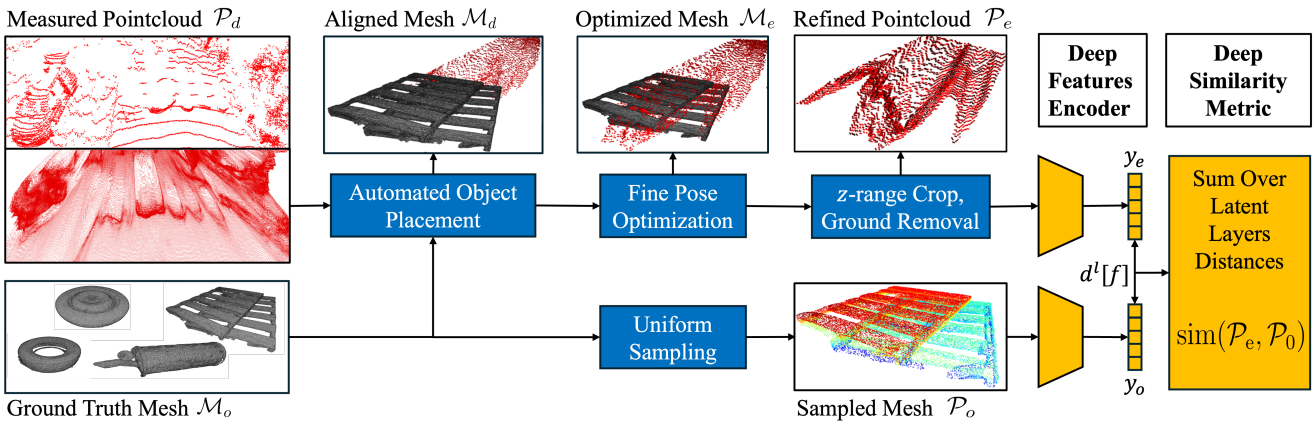

Figure 3. **Our 2T2S Pipeline.** Starting from a measured pointcloud $\mathcal{P}_d$, we reconstruct meshes for tiny, lost cargo objects and align them to the closest (15 or 25 meters) measured point cloud. We then perform a Fine Pose Optimization to reduce residual pose error and obtained the final, Refined Evaluation Pointcloud $\mathcal{P}_e$ by cropping noisy points belonging to ground. This cluster and the sampled ground-truth mesh $\mathcal{P}_o$ are subsequently fed into a feature encoder: the distances between the so obtained latent-layers vectors are normalized and summed to obtain our novel 2T2S deep similarity metric for robust perception evaluation of semantically and structurally incomplete objects.

Wasserstein-p distance W_p between corresponding column vectors:

$$d^{(l)}[f] = W_p\left(\tilde{Y}_{\mathrm{e}}^{(l)}[:, f],\ \tilde{Y}_0^{(l)}[:, f]\right), \tag{25}$$

This produces a distance vector $d^{(l)} \in \mathbb{R}^{F_l}$ for each latent layer. The distances are averaged across feature dimensions to obtain a scalar per layer:

$$\bar{d}^{(l)} = \frac{1}{F_l} \sum_{f=1}^{F_l} d^{(l)}[f]. \tag{26}$$

Finally, the similarity between $\mathcal{P}_{\mathrm{e}}$ and $\mathcal{P}_0$ is quantified as the sum over all latent layers:

$$\mathrm{sim}(\mathcal{P}_{\mathrm{e}}, \mathcal{P}_0) = \sum_{l=1}^{L} \bar{d}^{(l)}. \tag{27}$$

This approach enables robust comparison of object similarity even when geometry is incomplete or noisy.

5. Evaluation Setup

To validate our approach we generate 3D pointclouds by projecting 2D depth maps estimated with different modalities. For the gated-RCCB cross-spectral stereo setup we use Gated RCCB Stereo [12], which showed high quality lost-cargo depth estimation capabilities. For RCCB and RGB stereo, we use the widely adopted IGEV-Stereo [86]. Finally, we employ the monocular depth foundation model Metric3Dv2 [37] for the left monocular RGB camera. All depth estimation methods are finetuned on a training split for the summer and winter captures, supervising with a LiDAR ground truth generated projecting the Velodyne VLS-128 pointclouds into the corresponding images. Implementation details are provided in the Supplemental Material.

Reference Metrics. Depth-based metrics such as SiLog, RMSE, MAE, ARD, and Abs Rel are conventionally computed globally over an entire sensor-view. To focus on the regions of interest, we render the ground truth meshes $\mathcal{P}_0$ and $\mathcal{P}_e$ into aligned depth maps and apply the metrics locally within the object areas. In particular, each point $\mathbf{p}_i^o \in \mathbb{R}^3$ is mapped to screen space $\mathbf{p}_i^s \in \mathbb{R}^2$ following standard screen-space projection [65] as

$$\mathbf{p}_i^o \in \mathbb{R}^3 \xrightarrow{\text{proj}} \mathbf{p}_i^s \in \mathbb{R}^2,$$

and depth errors are computed only at pixels that are valid (non-zero) in both maps. Chamfer Distance and Voxel IoU are computed directly in 3D without projecting; for IoU we discretize space using a uniform voxel size s_{v} in all experiments. We apply the listed metrics on the lost cargo object point clouds at varying distances, and estimate the consistency of the changes in metric value as distance increases.

Metrics Application to Measurements. Summer and winter data was analyzed separately due to different measurement distances. For both datasets, metric values were averaged over all objects at each distance. This was computed individually for each perception method and yields a general estimation of how a metric value behaves when the distance of lost cargo objects increases.

The 2T2S metric was obtained according to Equation 26 from latent features of $\mathcal{P}_0$ and $\mathcal{P}_e$ at each measured distance. The metric value was used as the indicator of dissimilarity between ground truth and measurement. Of the reference metrics, Chamfer Distance and Voxel IoU were computed directly between P_e and P_0. The remaining reference metrics were computed on the rendered depth images as shown in the previous section. The 2T2S feature metric was obtained from SPVCNN [69] point cloud feature

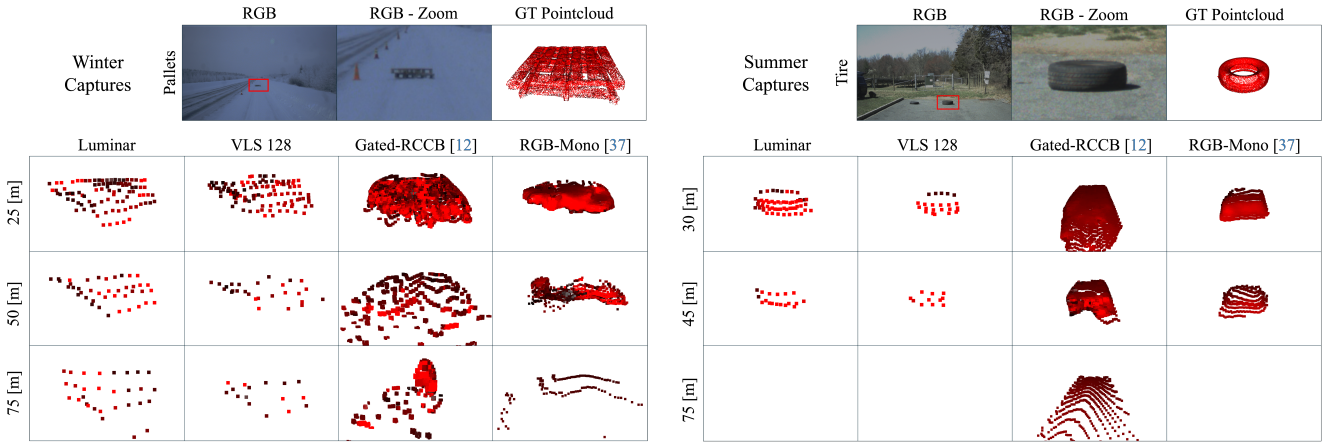

Figure 4. **Winter and Summer Reconstruction Examples.** Under varying weather and ranges, both LiDAR sensors (Luminar, Velodyne VLS-128) and learning-based depth estimators (stereo [12], monocular [37]) often struggle to recover the shape and distance of small lost-cargo objects. With LiDAR, even when only a few returns may hit the object, classic depth metrics can look favorable, even though geometry and semantics are poorly captured. Image-based stereo and monocular methods, in turn, often miss the same tiny objects at long distances due to reduced depth accuracy

encoder implemented as in [21] and trained on the ModelNet40 dataset [84]. The features are obtained from the first convolution layer of 'stage1', 'stage2', 'stage3' and 'stage4' encoder sequences of SPVCNN.

6. Results

Consistency Evaluation. We report results in Figure 5, which shows metric values of all methods across different evaluation distances. We seek a metric that captures perceptual degradation with distance and reflects the LiDAR–camera trade-off: LiDAR provides high depth accuracy but sparse sampling, whereas cameras offer dense geometry at lower absolute accuracy. Conventional metrics fail to reflect this trade-off, often producing misleading results. For instance, RMSE favors both Luminar and VLS-128 LiDARs point accuracy, as both score the best, while completely ignoring that their inherent sparsity prevents them from capturing object shapes, as seen qualitatively in Figure 4, where lost cargo objects can not be distinguished already at small ranges. All other common depth-based metrics indicate similar results and are presented in the Supplementary Material. Similarly, Chamfer Distance yields inconsistent rankings, with RCCB stereo [86] achieving on average best performances, but with high variation of best performing method across different (Summer and Winter) capturing conditions and ranges. Moreover, it does not clearly separate LiDAR from camera-based methods, limiting its discriminative power. The Voxel IoU generally ranks camera based methods above LiDAR based methods, with large outliers in the winter data, showcasing bias toward sheer point cloud density rather than accurately reflecting reconstruction quality. Our proposed 2T2S metric aligns with qualitative observations by favoring camera-based methods for their superior shape and semantic capture. It consistently ranks Gated RCCB Stereo [12] as the top performer (averaging 0.142 in summer and 0.137 in winter), while VLS-128 and Luminar sensors score the lowest due to their sparse coverage, which limits the representation of small-object geometry. This clear separation and stable ranking across all evaluation distances, unlike the fluctuating and counterintuitive results from other metrics, demonstrates that 2T2S provides a more discriminative and reliable measure of reconstruction quality for lost cargos.

Ablation studies. We conduct an ablation study using different neural network backbones to obtain the final 2T2S metric. We choose Point transformer V3 [83], SparseUNet [21], OA-CNN [59] and SPVCNN [69]. For each architecture, the feature vector for similarity comparison is taken from a suitable intermediate layer, as detailed in the Supplementary Material. Following the previous evaluation, Figure 6 shows average scores for summer and winter and across different ranges, with consistent behavior across architectures and differences largely limited to scale. This demonstrates that our method is not limited to SPVCNN outputs but generalizes across different point cloud encoders, highlighting its robustness and adaptability to newer, stronger architectures as they become available. We use SPVCNN in the main results as it provides the clearest separation between the tested methods and best matches qualitative observations (Fig. 4).

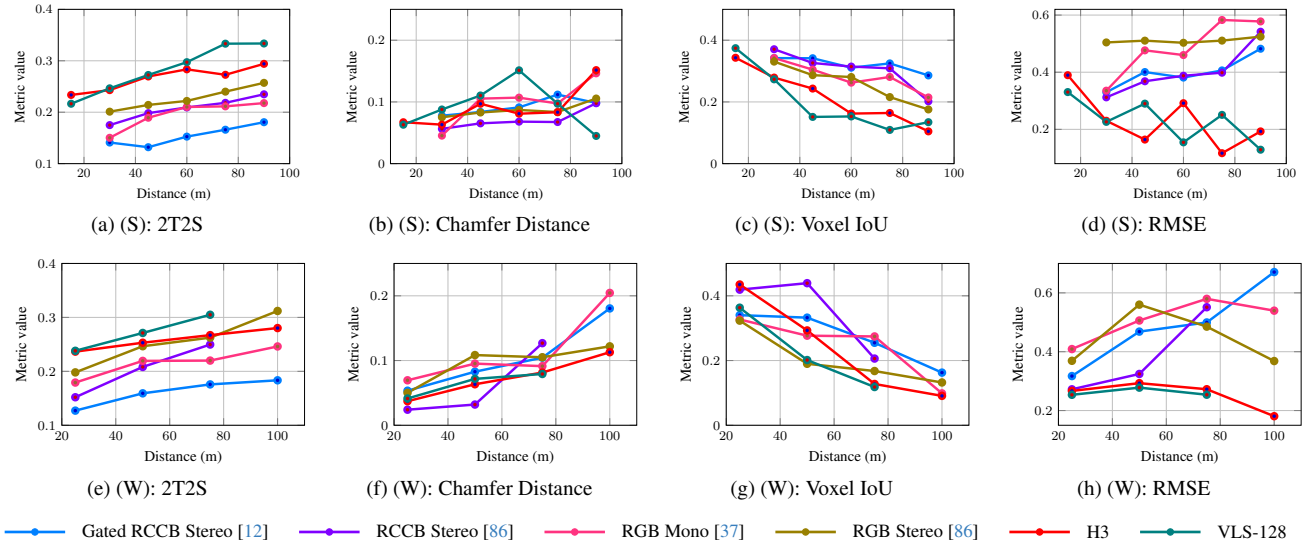

Figure 5. **Average Distance-Binned Evaluation Results.** We present averaged metric values by distance for each perception method on the two different captures of our dataset: winter (W) and summer (S) recordings. We note that the 2T2S metric values correlate with the qualitative results with smaller values (higher similarity to ground truth) for camera-based methods, with Gated RCCB Stereo standing out as the best performing method.

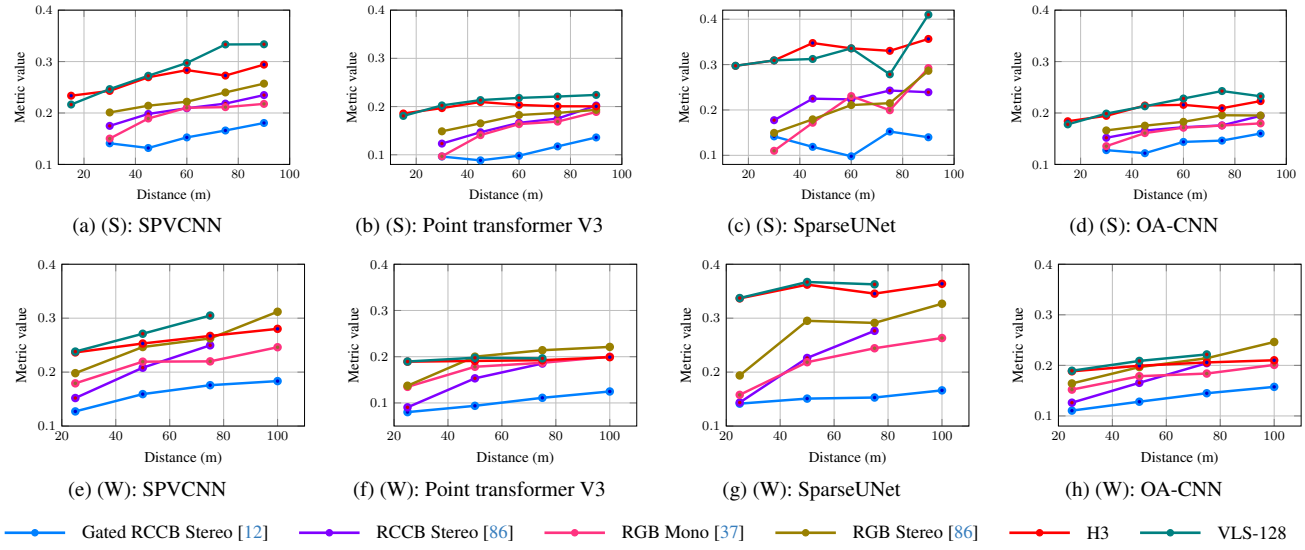

Figure 6. **Ablation Study.** Averaged metric values by distance for different feature encoder architectures in Winter (W) and Summer (S).

7. Conclusion

We introduced and release a novel dataset capturing very small and hard-to-find objects, to support research advances in the critical challenge of detecting obstacles in autonomous driving applications. To address the current limitations of existing evaluation metrics, that overlook object-level fidelity at distance, we complement it with an evaluation framework for long-range depth prediction of small objects. Specifically, our proposed 2T2S metric captures the LiDAR-camera trade-off, separates methods consistently, and quantifies object-level similarity between predicted and reference obstacle geometry in a learned point cloud feature space. Experiments across sensors, conditions, and encoder backbones demonstrate that 2T2S provides robust and modality-agnostic evaluation, offering a reliable tool for advancing depth perception research in safety-critical applications. Earlier reliable estimation of small obstacles increases effective detection range, which provides larger planning margins and can enable higher safe driving speeds under stopping distance constraints.

References

[1] Amit Adam, Christoph Dann, Omer Yair, Shai Mazor, and Sebastian Nowozin. Bayesian time-of-flight for realtime shape, illumination and albedo. 39(5):851–864, 2017. 2

[2] Evangelos Alexiou and Touradj Ebrahimi. Towards a point cloud structural similarity metric. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE Computer Society, 2020. 1, 2

[3] Pierre Andersson. Long-range three-dimensional imaging using range-gated laser radar images. 45(3):034301, 2006. 2

[4] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016. 2

[5] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval, 2014. 2

[6] Abhishek Badki, Alejandro Troccoli, Kihwan Kim, Jan Kautz, Pradeep Sen, and Orazio Gallo. Bi3D: Stereo depth estimation via binary classifications. In *arXiv preprint arXiv:2005.07274*, 2020. 2

[7] Harry G. Barrow, Jay M. Tenenbaum, Robert C. Bolles, and Helen C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 659–663. Morgan Kaufmann Publishers Inc., 1977. 2

[8] Fausto Bernardini, Joshua Mittleman, Holly Rushmeier, Cláudio Silva, and Gabriel Taubin. The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics*, 5(4):349–359, 2002. 4

[9] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14 (2):239–256, 1992. 5

[10] Mario Bijelic, Tobias Gruber, and Werner Ritter. A benchmark for lidar sensors in fog: Is detection breaking down? In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 760–767, 2018. 2

[11] Mario Bijelic, Tobias Gruber, and Werner Ritter. Benchmarking image sensors under adverse weather conditions for autonomous driving. 2018. 2

[12] Samuel Brucker, Stefanie Walz, Mario Bijelic, and Felix Heide. Cross-spectral gated-rgb stereo depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21654–21665, 2024. 1, 2, 3, 4, 6, 7, 8

[13] Jens Busck. Underwater 3-D optical imaging with a gated viewing laser radar. 2005. 2

[14] Jens Busck and Henning Heiselberg. Gated viewing and high-accuracy three-dimensional laser radar. 43(24):4705–10, 2004. 2

[15] J. J. Cabrera, A. Santo, A. Gil, C. Viegas, and L. Payá. Minkunext: Point cloud-based large-scale place recognition using 3d sparse convolutions, 2024. 2

[16] 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*, pages 11621–11631, 2020. 1

[17] A. Carballo, J. Lambert, A. Monrroy, D. Wong, P. Narksri, Y. Kitsukawa, E. Takeuchi, S. Kato, and K. Takeda. Libre: The multiple 3d lidar dataset. In *IEEE Intelligent Vehicles Symposium (IV)*, 2020. 2

[18] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 2

[19] Jagpreet Chawla, Nikhil Thakurdesai, Anuj Godase, Md Reza, David Crandall, and Soon-Heung Jung. Error diagnosis of deep monocular depth estimation models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5344–5649. IEEE, 2021. 1

[20] Jaesung Choe, Kyungdon Joo, Tooba Imtiaz, and In So Kweon. Volumetric propagation network: Stereo-lidar fusion for long-range depth estimation. *IEEE Robotics and Automation Letters*, 6(3):4672–4679, 2021. 2

[21] Pointcept Contributors. Pointcept: A codebase for point cloud perception research. https://github.com/Pointcept/Pointcept, 2023. 7

[22] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017. 2

[23] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 1, 2

[24] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, pages 2366–2374, 2014. 1, 2

[25] Yu Fan, Zicheng Zhang, Wei Sun, Xiongkuo Min, Ning Liu, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai. A no-reference quality assessment metric for point cloud based on captured video sequences. In *2022 IEEE 24th international workshop on Multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2022. 1, 2

[26] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023. 1, 2

[27] Ravi Garg, B.G. Vijay Kumar, Gustavo Carneiro, and Ian Reid. Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756, 2016. 2

[28] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1

[29] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference*

on computer vision and pattern recognition, pages 270–279, 2017. 2

[30] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 1, 2

[31] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[32] Jiaqi Gu, Zhiyu Xiang, Yuwen Ye, and Lingxuan Wang. Denselidar: A real-time pseudo dense depth guided depth completion network. *IEEE Robotics and Automation Letters*, 6(2):1808–1815, 2021. 1

[33] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2485–2494, 2020. 2

[34] Akhil Gurram and Antonio M. López. On the metrics for evaluating monocular depth estimation. *arXiv preprint arXiv:2302.10007*, 2023. 2

[35] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012. 2

[36] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Towards precise and efficient image guided depth completion. 2021. 2

[37] 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. 1, 2, 6, 7, 8

[38] Jiwen Jia, Junhua Kang, Lin Chen, Xiang Gao, Borui Zhang, and Guijun Yang. A comprehensive evaluation of monocular depth estimation methods in low-altitude forest environment. *Remote Sensing*, 17(4):717, 2025. 1

[39] Maria Jokela, Matti Kutila, and Pasi Pyykönen. Testing and validation of automotive point-cloud sensors in adverse weather conditions. *Applied Sciences*, 9, 2019. 2

[40] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1

[41] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. *Computer Vision and Image Understanding*, 191: 102877, 2020. 1

[42] Andreas Kolb, Erhardt Barth, Reinhard Koch, and Rasmus Larsen. Time-of-flight cameras in computer graphics. In *Computer Graphics Forum*, pages 141–159. Wiley Online Library, 2010. 2

[43] Robert Lange. 3D time-of-flight distance measurement with custom solid-state image sensors in CMOS/CCD-technology. 2000. 2

[44] Martin Laurenzis, Frank Christnacher, and David Monnin. Long-range three-dimensional active imaging with superresolution depth mapping. 32(21):3146–8, 2007. 2

[45] Martin Laurenzis, Frank Christnacher, Nicolas Metzger, Emmanuel Bacher, and Ingo Zielenski. Three-dimensional range-gated imaging at infrared wavelengths with super-resolution depth mapping. In *SPIE Infrared Technology and Applications XXXV*, 2009. 2

[46] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 1

[47] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical Stereo Matching via Cascaded Recurrent Network with Adaptive Correlation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16242–16251, New Orleans, LA, USA, 2022. IEEE. 1

[48] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241, 2021. 2

[49] Zhaoshuo Li, Xingtong Liu, Nathan Drenkow, Andy Ding, Francis X. Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6177–6186, 2021. 2

[50] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022. 2

[51] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021. 1

[52] Yongfan Liu and Hyoukjun Kwon. Efficient depth estimation for unstable stereo camera systems on ar glasses. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6252–6261, 2025. 1

[53] Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David Gealy, and Ken Goldberg. Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In *2018 IEEE International Conference on robotics and automation (ICRA)*, pages 5620–5627. IEEE, 2018. 1

[54] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning, 2018. 1

[55] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18237–18246, 2023. 1

[56] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cam-

eras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 1

[57] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 1

[58] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 2

[59] Bohao Peng, Xiaoyang Wu, Li Jiang, Yukang Chen, Hengshuang Zhao, Zhuotao Tian, and Jiaya Jia. Oa-cnns: Omni-adaptive sparse cnns for 3d semantic segmentation, 2024. 7

[60] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[61] Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, Mihai Dolha, and Michael Beetz. Towards 3d point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927–941, 2008. 4

[62] Michael Schober, Amit Adam, Omer Yair, Shai Mazor, and Sebastian Nowozin. Dynamic time-of-flight. In *CVPR*, pages 6109–6118, 2017. 2

[63] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE, 2015. 2

[64] Brent Schwarz. Lidar: Mapping the world in 3D. *Nature Photonics*, 4(7):429, 2010. 1, 2

[65] Peter Shirley, Michael Ashikhmin, and Steve Marschner. *Fundamentals of computer graphics*. AK Peters/CRC Press, 2009. 6

[66] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 1

[67] 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

[68] Lior Talker, Aviad Cohen, Erez Yosef, Alexandra Dana, and Michael Dinerstein. Mind the edge: Refining depth edges in sparsely-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10606–10616, 2024. 1

[69] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution, 2020. 6, 7

[70] Jiexiong Tang, John Folkesson, and Patric Jensfelt. Sparse2dense: From direct sparse odometry to dense 3-d reconstruction. *IEEE Robotics and Automation Letters*, 4(2): 530–537, 2019. 2

[71] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing*, 30:1116–1129, 2020. 2

[72] Junzhe Zhang Tai WANG Ziwei Liu Dahua Lin Tong Wu, Liang Pan. Density-aware chamfer distance as a comprehensive metric for point cloud completion. In *In Advances in Neural Information Processing Systems (NeurIPS), 2021*, 2021. 2

[73] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 international conference on 3D Vision (3DV)*, pages 11–20. IEEE, 2017. 1, 2

[74] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition, 2018. 2

[75] Amanpreet Walia, Stefanie Walz, Mario Bijelic, Fahim Mannan, Frank Julca-Aguilar, Michael Langer, Werner Ritter, and Felix Heide. Gated2gated: Self-supervised depth estimation from gated images. 2022. 2

[76] Celyn Walters, Oscar Mendez, Mark Johnson, and Richard Bowden. There and Back Again: Self-supervised Multispectral Correspondence Estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5147–5154, 2021. 1

[77] Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, and Felix Heide. Gated stereo: Joint depth estimation from gated and wide-baseline active stereo cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13252–13262, 2023. 1, 2

[78] Jialiang Wang, Daniel Scharstein, Akash Bapat, Kevin Blackburn-Matzen, Matthew Yu, Jonathan Lehman, Suhib Alsisan, Yanghan Wang, Sam Tsai, Jan-Michael Frahm, Zijian He, Peter Vajda, Michael F. Cohen, and Matt Uyttendaele. A practical stereo depth system for smart glasses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21498–21507, 2023. 1

[79] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1

[80] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 1

[81] Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5249–5260, 2025. 1

[82] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 12747–12756, 2021. 2

[83] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger, 2024. 7

[84] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 7

[85] Wang Xinwei, Li Youfu, and Zhou Yan. Triangular-range-intensity profile spatial-correlation method for 3D super-resolution range-gated imaging. 52(30):7399–406, 2013. 2

[86] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21919–21928, 2023. 1, 6, 7, 8

[87] Zhenyu Xu, Yuehua Li, Shiqiang Zhu, and Yuxiang Sun. Expanding sparse lidar depth and guiding stereo matching for robust dense depth estimation. *IEEE Robotics and Automation Letters*, 8(3):1479–1486, 2023. 1

[88] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[89] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1

[90] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1, 2

[91] Yongjian Zhang, Longguang Wang, Kunhong Li, Zhiheng Fu, and Yulan Guo. Slfnet: A stereo and lidar fusion network for depth completion. *IEEE Robotics and Automation Letters*, 7(4):10605–10612, 2022. 1, 2

[92] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18527–18536, 2023. 1

[93] Tiancheng Zhi, Bernardo R Pires, Martial Hebert, and Srinivasa G Narasimhan. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1916–1925, 2018. 1

[94] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2