Toward open benchmark tests for automotive lidars, year 1: static range error, accuracy, and precision

Zach Jeffries,[a] **Jeremy P. Bos**,[a,*] **Paul McManamon**,[b,*] **Charles Kershner**,[c] **and Akhil Kurup**[a]

[a]Michigan Technological University, Robust Autonomous Systems Laboratory, Department of Electrical and Computer Engineering, Houghton, Michigan, United States
[b]Exciting Technology, LLC, Dayton, Ohio, United States
[c]National Geospatial-Intelligence Agency (NGA), Springfield, Virginia, United States

Abstract. This paper describes the initial results from the first of 3 years of planned testing aimed at developing methods, metrics, and targets necessary to develop standardized tests for these instruments. Here, we evaluate range error accuracy and precision for eight automotive grade lidars; a survey grade lidar is used as a reference. These lidars are tasked with detecting a static, child-sized, target at ranges between 5 and 200 m. Our target, calibrated to 10% reflectivity and Lambertian, is a unique feature of this test. We find that lidar range precision is in line with the values reported by each manufacturer. However, we find that maximum range and target detection can be negatively affected by presence of an adjacent strong reflector. Finally, we observe that design trade-offs made by each manufacturer lead to important performance differences that can be quantified by tests such as the ones proposed here. This paper also includes some lessons learned, planned improvements, and discussion of future iterations of this activity. [DOI: 10.1117/1.OE.62.3.031211]

Keywords: LIDAR; LADAR; automotive; autonomous vehicles; autonomous perception.

Paper 20221061SS received Sep. 15, 2022; accepted for publication Dec. 13, 2022; published online Jan. 5, 2023.

1 Introduction

The options available to automobile manufacturers and Tier 1 integrators for low-cost light detection and ranging (LIDAR or lidar) sensors used for autonomous vehicle (AV) and advanced driver assistance systems (ADAS) applications are growing rapidly. Fundamentally, these lidars all use lasers to measure distance across a set field of view (FoV). Lidar engineers make design trade-offs to gain competitive advantages in performance and cost in what is a rapidly growing, highly competitive market. Some of these trade-offs include operating wavelength (typically between 850 and 1600 nm), range measurement based on either direct detection/Time-of-Flight (ToF) or coherent techniques, beam steering solutions (mechanically rotating components, MEMS mirrors, microlenses), and laser source type [vertical cavity surface-emitting lasers (VCSELs), edge-emitting diodes]. These design choices have trade-offs of their own, with differences in scan patterns, sampling frequency, achievable ranges, susceptibility to interference from other lidars, etc.

Lidar is one of the most important and versatile components of an AV's perception system. These sensors provide the vehicle with a three-dimensional (3D) map of the location of objects around the vehicle in all lighting conditions. They also allow estimation of the vehicle's position with respect to its surrounding, all updated hundreds of times per second. The performance of the complete AV system can be directly linked to the collective performance of all components and subsystems.[1] AV system integrators must consider the likely performance variation of each component and the impact of that variation at the system and subsystem level.

On datasheets and sales, literature vendors list specifications that assist engineers in designing AV perception systems. Some of these specifications, FoV and angular sampling rate, e.g.,

*Address all correspondence to Jeremy P. Bos, jpbos@mtu.edu; Paul McManamon, paul@excitingtechnology.com

are easy to verify. Others are more ambiguous. For example, range is often listed but is only occasionally accompanied by a target reflectivity. Similarly, range precision is often listed but usually as a normal deviation from the mean; it is unclear if this is a reliable assumption. In comparing one lidar to others, integrators often resort to evaluation of engineering samples. This is a costly endeavor and risks introducing bias due to experiment design, etc. Because these results are kept confidential and internal to those performing the test it is possible for two integrators to draw different conclusions about the same lidar unit. Both the need for a large-scale benchmarking activity and some attempts at testing standardization are clear.

While most lidar testing and benchmarking are confidential, there are some works available in the open literature. Glennie and Lichti[2] assessed a Velodyne HDL-64E and developed a calibration routine for mapping applications. Later, Glennie and Hartzel[3] compared a Livox Mid-40 and an Ouster OS1-64 to their published specifications. In Mittet et al.,[4] the range accuracy and precision of an early Quanergy M8 unit was examined. Kutila et al.[5] examined the effects of arctic conditions on lidar sensors. This work is both qualitative and quantitative comparing intensity and reported range from lidar scans in winter conditions to the average across multiple lidar sensors; it does not include a ground truth. Other work by the group including Rosenberg et al.[6,7] have focused on developing sensors models for use in simulation and necessarily involves characterizing sensors using metrics and benchmarks.

Attempts at codifying test methods and comprehensive benchmarking activities are a recent development. Cattini et al.[8] proposed a very precise laboratory method. Their procedure is too cumbersome and complex to be used in a field test event such as ours involving potentially dozens of lidars over a single day. However, their findings with respect to unit-to-unit variation, warm-up time, and stability will be useful as standards are developed. An extensive indoor test was performed in Ref. 9 involving ten rotary style lidars. This testing includes a ground truth and examines accuracy, precision, and intensity variations for a static target containing three materials. However, the specific reflectance behavior of these materials over angle and wavelength are not characterized. In addition, this work presents results in terms of the mean error and standard deviation from the mean. Kim et al.,[10] perform a competent evaluation of lidar performance under degraded conditions using a two-way ANOVA test. Their work focuses on number of points and intensity though rather than range accuracy and precision and involve only a single lidar. Test repeatability will inevitably be part of this activity as work progresses.

Occurring in parallel to this effort, the work by Schulte-Tigges et al.[11] evaluates six non-rotary automotive lidar devices against different static and dynamic targets. In addition, metrics and processing steps are outlined for each scenario. Some of which include a target detection algorithm. All the targets are relatively large and, like those in the work by Lambert et al.,[9] are not well characterized. Similarly, results are presented in terms of the deviation about the mean. Also, the measurement references are either hand-measured or derived from GPS.

By way of standards, the National Institute for Standards and Technology (NIST) and ASTM International have published standards for metric assessments of laser-based, scanning, ToF, single detector 3D imaging systems,[12,13] but no manufacturer advertises that their internal testing and calibrations are done to these standards, nor is there a requirement for them.

In this work, we describe the results of the first year of a proposed 3-year lidar benchmarking exercise. This effort began in 2019 with the intent of an initial public test at the SPIE Defense and Commercial Sensing conference in April of 2020. This event was cancelled due to the worldwide SARS-CoV-2 pandemic and finally reconvened at SPIE Defense and Commercial Sensing in April of 2022.

The aim of the first year of testing was evaluation of range accuracy and precision on static targets with the goal of refining processing, data acquisition, and test setup. The aim of years two and three is to add additional effects like oncoming lidars on the test range, weather, and dynamic targets. A complete description of proposed future efforts is found in the Appendices to this paper.

Our work differs from the previous and concurrent works in several important ways. First, this effort compares results between both rotary and scanning units. Most crucially, it also includes a small, child-sized, target calibrated to 10% reflectivity over the range between 800 and 1600 nm. This target, provided by Labsphere, is also verified as being a purely Lambertian reflector. Our year one testing also involves two configurations: a control configuration and a second identical

test setup with highly reflective adjacent objects we call "confusers." Unlike other previous works, we present results in terms of the median sample range and the sample interquartile range (IQR) to avoid the presumption that sample detections can be fit to a normal distribution. The effort described here also does not include any detection algorithms or weather effects though both are likely to be considered in future iterations of this event.

Testing was conducted in an open field in Kissimmee, Florida, among eleven different lidars including three of the same make and model. We present here a comparison between eight of the units tested using a survey-grade lidar as a reference. All the units evaluated have advertised maximum detection ranges of between 100 and 200 m and operating wavelengths between 800 and 950 nm. In our test configuration, the average maximum detection range was 50 m with the minimum of 25 m and a maximum of 120 m. Range accuracy across units was biased short of the reference distance by −0.6 cm excluding outliers. Range precision across all units is estimated as 3.6 cm compared with a typical advertised value of 3 cm for most units. RMS planar fitting errors with respect to the target differed by around 7 cm. The addition of adjacent "confusers" reduces range precision of all units by 25% to 65% depending on the metric. Confusers also significantly reduce the ability of all lidars to score targets at range. For this reason, the "maximum range" self-reported by lidar manufacturers should be considered measured under the best possible conditions. A finding of this work is that the presence of adjacent, highly reflective, object to a dim object reduces the probability of detecting the dim object. For this reason, this condition should be considered in the development of any standard.

Over the remainder of this paper, we describe our test setup and method starting in the next section. In the section that follows we describe our test metrics and processing approach. Results are presented in Sec. 3. Conclusions and directions for future work are provided in Sec. 4. In an appendix to this paper, we outline plans for years two and three of the effort and solicit both feedback and participation.

2 Test Setup

2.1 *Location*

Testing was conducted near the Bridging the Innovation Development Gap (BRIDG) center in Kissimmee, Florida (28.291525N, −81.371776 W) on April second and third, 2022. A dry run was conducted on the second, when the conditions were overcast with scattered thunderstorms, with temperatures between 61°F and 81°F. Relative humidity was between 79% and 99%. Testing was conducted on the third, conditions were sunny with temperatures between 73°F and 77°F and relative humidity between 56% and 60%.

The test area was comprised of the unimproved lot behind the BRIDG facility and Skywater Technologies buildings. An overhead plan view of the test area can be found in Fig. 1. The buildings and courtyard are seen in the lower half of the image. Neocities Way is directly behind the "origin" located just off the adjacent sidewalk. The test area itself is somewhat flat along the major axis of the target field. Black areas in the image indicate shadowed areas and suggest a depression present near the center of the test area and a decline in elevation to the right of the test area (top of the image). The device test location is labelled as "origin" at the center left of the image. Targets are labeled by their approximate distance from the origin with the far 200 m target located at the center right. The line between the 200 m target and the origin makes up the main axis of the test range. The image in Fig. 1 was generated using the Reigl VZ-400i by placing the unit at various locations around the test area. Each black circle in the image is the origin of one scan area. Data from multiple scan areas were combined to create the composite.

2.2 *Targets*

Both calibrated and other objects were present in the test range. As part of this benchmarking effort Labsphere developed a 15 cm × 80 cm flat aluminium target with a Lambertian coating that is 10% reflective from 800 nm to 1600 nm. The size corresponds, approximately, to the cross-section area of a small child when viewed from the side (Fig. 2). Only range data for child

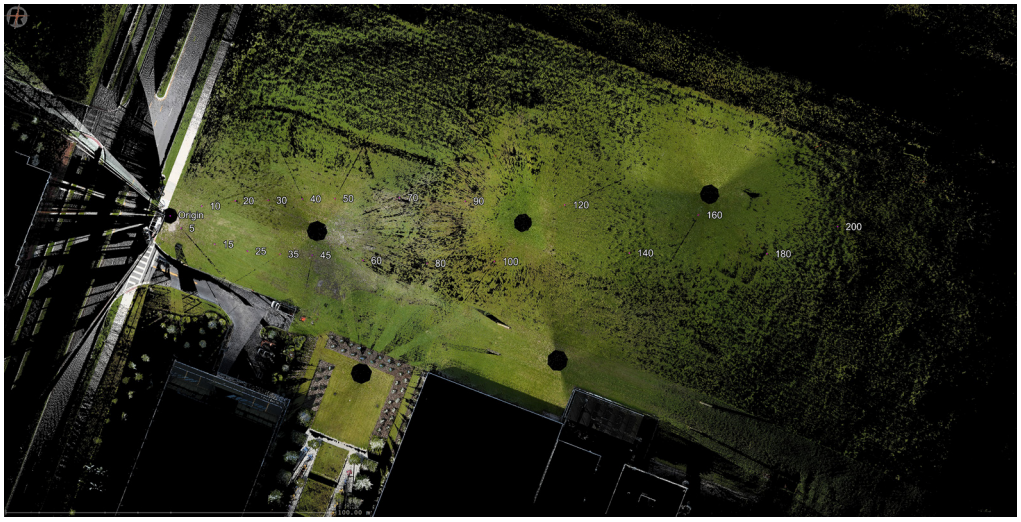

Fig. 1 Overview of the test area as a RGB point cloud assembled from multiple positions across the test range indicated by black circles. Black pixels indicate no return and are obstructed or shadowed due to observation geometry.

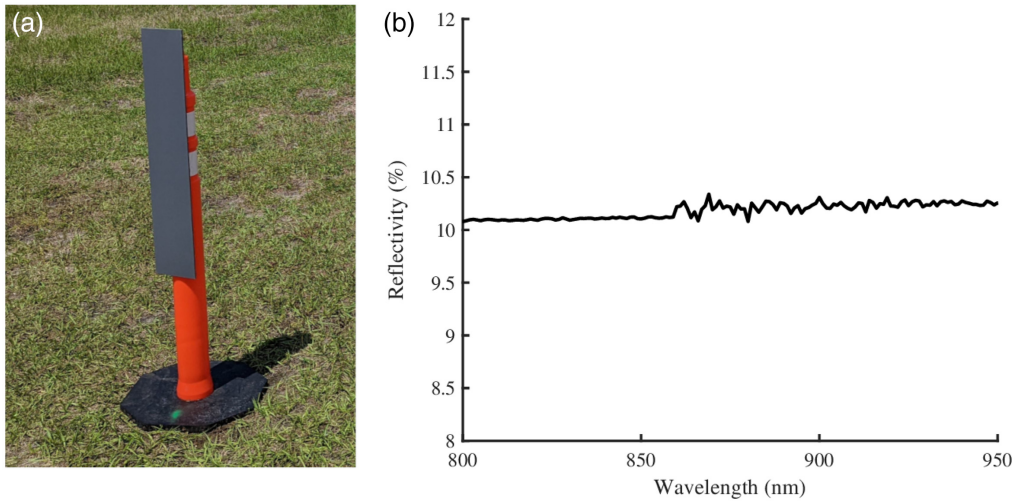

Fig. 2 (a) Labsphere child sized target (gray) affixed to delineator traffic cone with attached base Targets are 10% reflective at lidar operating wavelengths and Lambertian scattering. (b) Target reflectivity as measured by Labsphere over the range of operating wavelengths for the lidars tested.

size targets are presented in this work. Figure 2 also includes a plot of the measured reflectivity of the target between 800 and 950 nm, the range of operating wavelengths for the lidars tested. The exact mean reflectance over this range is 10.18% with a minimum value of 10.08% and a maximum of 10.34%.

All monostatic lidars, no matter their technology, rely on light emitted from a common source location and received at the same location. Whether scanning or flash illuminated the received power per solid angle from a target decreases with range. Naturally, also, the area occupied by the solid angle sampled by the lidar also increases with distance. For this reason, in a test like this one the targets must be arranged so that they do not overlap. This is straightforward for turntable scanning type lidars assuming there is enough open area around the device under test (DUT). However, an increasing number of automotive lidars have a limited azimuthal FoV. For this reason, the test design aims to include the maximum number of targets within a 60°FoV.

Targets were arranged starting with the 200 m target aligned along the intended optical axis down the test range. Starting from the origin the first two targets were placed in 5 m increments

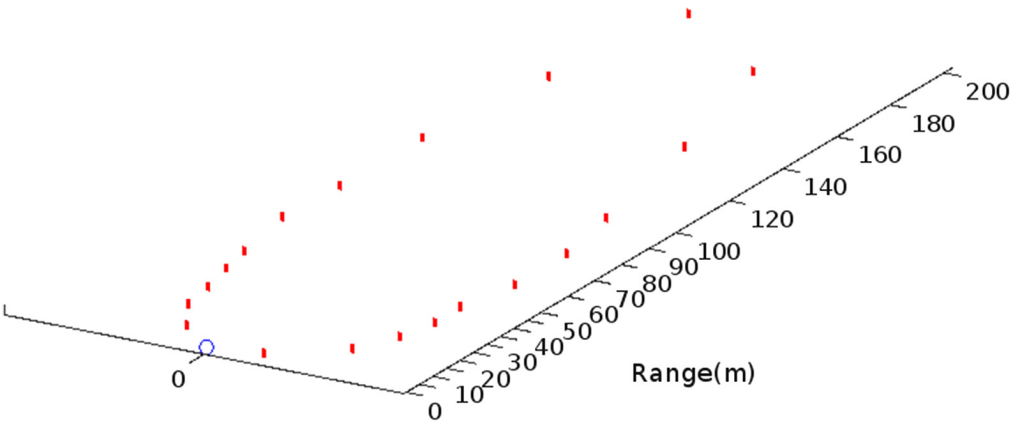

Fig. 3 Nominal plan view of the test range. The long axis of the range is oriented between the test location at "0" and the last target at 200 m. Targets inside 100 m are oriented along an ~60 deg FoV.

alternating along the right and then left side of a ±30 deg FoV with respect to the text axis. Additional targets were placed in the same manner in 5 m increments out to 50 m on the left side of the range. The next set of targets were placed in 10 m increments starting on the right side of the range and continuing out to 100 m. At 100 m, the spacing was increased to 20 m out to the final target at 200 m. An idealized plan layout of this configuration can be found in Fig. 3.

Each Labsphere target was fixed to a delineator-type traffic cone with stabilizing base using self-adhesive hook and loop faster. Each target was aligned such that the center of the target was approximately aligned with the horizontal optical axis. This alignment was done manually using a spotting scope from test origin. The nature of the cone and base used to hold the targets means that the target itself is at best orthogonal to the ground patch on which the cone is placed, but it was not possible to ensure that the target itself is orthogonal to the ray between the target center and the origin. For this reason, targets will not be oriented uniformly with respect to the origin. However, each DUT observes each target in the same geometry. Also, using our reference system, we were able to measure the orientation of each target. This information is summarized in Table 5 in the appendix. The limited vertical angular FoV of the test targets is a known weakness of this test setup and will be addressed in subsequent efforts.

Testing was performed in two configurations referred to as "Lane 1" and "Lane 2." Lane 1 consisted of only the targets affixed to the stands. Lane 2 consisted of the targets intermixed with "real-world" objects as scene clutter or "confusers." These objects included: orange and white folding metal traffic barricades with retroreflective panels, orange rubber traffic cones with retroreflective tape, and orange plastic delineator tubes with retroreflective tape along the left side of the lane, and a variety of 48-in. steel traffic control signs containing black text on a retroreflective orange background along the right side (signs consisted of type II road construction signs; one "reverse curve" sign (MUTCD code W1-4L), two "two way traffic signs" (W6-3), two "one lane road ahead" (W20-4), and two "be prepared to stop" signs (W3-4)[14]). These confusers were placed adjacent to each target. This arrangement provides the opportunity to test the effect of laser power automatic gain control and its impact on range detection performance. Figure 4 shows the test set up of Lane 2, using confusers as well as the test targets. The location of the test targets was the same in both configurations.

2.3 *Instruments*

2.3.1 *Reference lidar*

A survey-grade Riegl VZ-400i Terrestrial Lidar Scanner (TLS) collected the high-resolution, high-accuracy point clouds used as our reference data. Lidar scanners of this type and accuracy have been used to collect reference data for similar efforts.[3] Table 1 contains the specifications for this instrument.

Fig. 4 Photograph of Lane 2, including adjacent reflective "confusers," from the perspective of the test origin.

Table 1 Riegl VZ-400i specifications.[15]

Max measurement range ($\rho \geq 20\%$)	120 m
Max measurement range ($\rho \geq 90\%$)	250 m
Accuracy	5 mm
Precision	3 mm
Beam divergence	0.35 mrad
Max. targets per pulse	4
Laser wavelength	1550 nm

The TLS was mounted on a leveling tribrach with a removable insert, which was secured atop a sturdy survey tripod. Each scan was collected using 0.02 deg horizontal and vertical angular sampling, at a scan rate of 1.2 mHz. A camera affixed to the top of the scanner collected color images that were used to apply and RGB value to each point in the point cloud. The resulting data had an average of 300 points/m^2 on the target features. Two complete reference datasets were collected, one for each lane configuration. Each dataset comprised of multiple scans collected at different locations along the test lanes, ensuring complete, high density coverage of the area, ground, targets, and buildings (see Fig. 1).

The top of the tribrach was set ~1 m off the ground and leveled. After a reference scan, the TLS was removed from the tribrach so each test lidar could be affixed to the tribrach using the removable inserts. This setup ensured that each test lidar's coordinate system and pose could easily be aligned to the project's reference coordinate system using the mounting point offsets supplied in the reference documentation.

One scan in each reference dataset was used to define the coordinate system for each test lane configuration. In processing the reference data, a local Cartesian coordinate system was defined such that the origin corresponds to the TLS's X-Y origin at the scan position at the end of the test lane and the Z origin corresponding to the top of the tribrach, calculated by subtracting the optical center from the base plate of the instrument. The elevation from the ground of the optical axis for the DUTs would be slightly lower than the TLS at 92 cm.

2.3.2 *Test lidars*

Data were collected for each lane configuration using eight different lidar devices. Three of the lidars were of the same make and model. Since our objective is to evaluate variation between

lidar designs and not individual lidar performance the make and model are obscured. Instead, we refer to each device with an assigned letter between "A" and "H." Some general observations regarding the DUTs: all of the test lidars operated near between 800 and 950 nm. The test pool consisted of nearly equal portion of MEMs or other static scanning lidars and traditional rotating scanning devices. Most DUTs indicated a range precision near 3 cm and an operating range between 100 and 200 m.

Each DUT was aligned to the optical axis at the origin of the test setup. DUTs were connected to a laptop computer running Ubuntu 18.04 and Robot Operating System (ROS) Melodic.[16] DUTs were configured using the default settings for the respective ROS driver. Using the rosbag tool, 100 consecutive "pointcloud2" messages were collected from each DUT in each lane configuration.

2.4 *Data Processing*

Data from the Riegl TLS was processed using the manufacturer's software (RiScan Pro 2.14.1). Individual scans were automatically registered together on board the scanner during collection and further refined using multistation adjustment, a plane-fitting registration routine in RiScan Pro. Points with very low intensity or high pulse deviation were then filtered out, and RGB information was added to each point.

Each DUT point cloud was exported from ROS to MATLAB and initially registered to the Riegl reference point cloud manually. Final alignment was completed using iterative closest point (ICP) matching. For each DUT, the 100 "pointcloud2" messages were combined into a single point cloud data object and then aligned to the Riegl reference point-cloud via ICP. The result of this pre-processing step is a best-effort aligned point cloud that uses a single transform between the DUT coordinate system and reference system with all 100 point-cloud messages combined into a single data object.

Data was then further processed for each range target. For each candidate sensor, each target was identified and captured by bounding boxes first from a top-down point-of-view, then a left-facing point-of-view and finally a forward-facing point-of-view. Special care was taken to include neither the cones that held the targets nor confusers when present. This manual process was repeated until every target scored by each sensor was identified.

2.5 *Scoring and Metrics*

Figure 5 shows the point clouds collected and the points labeled as the 10 m target with only the initial registration. Scoring was performed by finding the minimum distance from each lidar

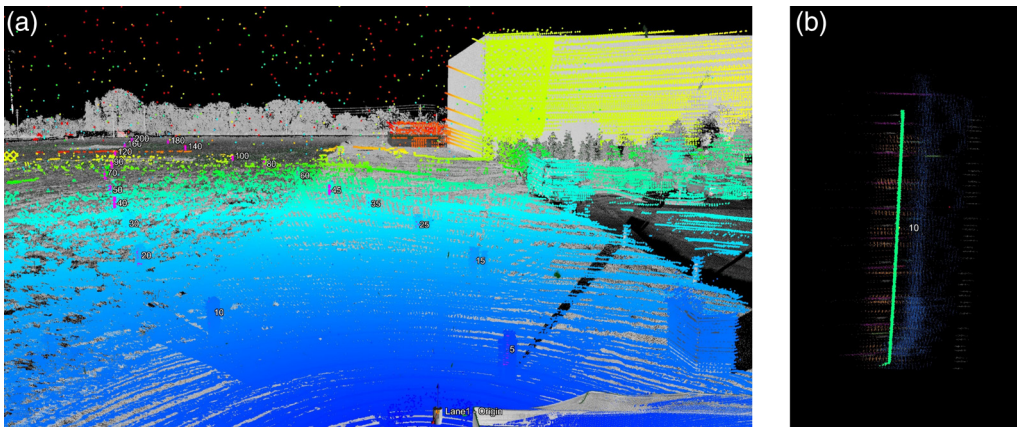

Fig. 5 (a) The reference point cloud scan (gray) overlayed with point clouds collected by each of the DUT lidars (colors). (b) Side view of an initial alignment between the reference point cloud (green) and point clouds from the DUT lidars for the 10 m target. Notice that the target is tilted toward the test origin.

point reported by the DUT to the closest coplanar point from the reference point cloud considering only range and cross-range dimensions; difference in elevation is not considered.

For this initial examination, only range statistics are examined: range accuracy and precision provided by each DUT with respect to the reference. The RMS plane fit is also provided and refers to the total fit RMS error between each reported DUT target point and the nearest reference point on the target excluding elevation.

3 Results

Results are presented here with as little interpretation as possible apart from highlighting cases where the field test results further motivate additional testing and the development of standard. Herein, we consider a target to be detected only if 20 or more points over the 100 ROS point clouds score the target. Also, some results are provided out to the maximum scored range the number of points on target beyond 50 m is typically <200 points per target implying <2 points per scan on average. For this reason, results beyond 50 m should be considered informative rather than descriptive.

In terms of overall performance, Lidar B was scored out to 120 m in both lanes. Though the DUT was not scored on the 80 and 100 m targets in Lane 2 with confusers present. Lidar H was scored out to 90 m in Lane 1 but on 45 m in Lane 2. Similarly, DUTs C and F were scored out to 50 m in Lane 1 but only to 45 m in Lane 2. DUTs D and E scored to 40 m in Lane 1 and to 30 and 25 m in Lane 2, respectively. Lidar A scored to 30 m in both lanes and lidar G to 25 m also in both lanes. Only targets between 10 and 25 m are scored by all DUTs.

Tabulated results for the eight DUTs are provided in Tables 2 and 3 representing the results with and without the confusers. Beginning with the top row we observe that three of the eight units reported the targets closer to the origin compared to the reference. Across all units, the absolute average error in position estimate was 2.9 cm and the minimum to maximum variation (span) was 12.4 cm. Adding confusers increases the absolute average error across all DUTs and targets to 4.8 cm; a 65% increase. With the confusers, the span also increased to 15.4 cm (25%). RMS plane fit error increased from 7.4 to 9.6 cm an increase of 30%.

Range precision averaged 3.6 cm across all targets and DUTs and 3.1 cm excluding lidar F, in line with the typical advertised value of 3 cm. However, on a target-by-target and device-by-device basis, there is quite a bit of discrepancy from a minimum variation of 0.7 cm for lidar C on the 5 m target to 15.2 cm for lidar H on the 90 m target. While it may be assumed this is simply an effect of range, the variation of lidar D observing the 35 m target was 14.9 cm. Adding confusers increases the average range ambiguity (decreases precision) across the test population by 26% to 4.6 cm. The minimum range precision remains with lidar C on the 5 m target while lidar's D precision observing the 35 m target increases to 17 cm.

Table 2 Results for Lane 1 or the configuration without confusers.

Metric/DUT	A	B	C	D	E	F	G	H
Mean range accuracy (cm)	−4.11	−4.21	0.61	1.64	1.01	8.16	−1.28	2.09
Worst target accuracy (cm)	−6.5	−17.33	2.95	4.91	−2.84	23.25	−3.95	15.84
Best target accuracy (cm)	−1.44	0.26	0.27	0.69	0.82	−0.12	−0.02	0.49
RMS plane-fit error (cm)	6.43	6.63	3.23	8.3	4.56	10.95	3.36	15.41
Range precision (cm)	2.92	3.45	1.33	7.57	3.07	4.7	1.8	4.29
Worst target precision (cm)	5.23	7.19	3.18	14.87	4.48	7.2	2.76	15.24
Best target precision (cm)	3.15	0.72	0.64	2.97	1.52	1.45	0.9	2.54
Total points on target (all)	25,037	34,409	82,543	29,696	6677	82,167	6787	33,608

Table 3 Results for Lane 2, the configuration with confusers.

Metric/DIT	A	B	C	D	E	F	G	H
Mean range accuracy (cm)	−1.93	−3.95	−4.96	2.5	−6.96	8.42	−3.11	6.36
Worst target accuracy (cm)	6.62	−16.84	−9.05	17.71	20.08	18.41	−8.68	11.07
Best target accuracy (cm)	0.91	0.6	−0.73	1.02	1.11	1.29	0.45	0.35
RMS plane-fit error (cm)	6.89	11.31	6.38	9.79	13.05	14.02	4.2	11.1
Range precision (cm)	4.42	5.51	2.22	7.57	5.18	6.59	1.82	3.61
Worst target precision (cm)	7.91	9.56	5.82	17.4	11.97	11.67	2.87	5.53
Best target precision (cm)	3.15	2.58	0.68	2.72	2.4	3.07	0.76	2.69
Total points on target (all)	19,208	27,633	79,570	23,675	5077	70,159	5716	8757

Tables 2 and 3 also report the number of points across all targets. From this information, the reader can glean some information about the relative sampling rate in the angle space occupied by the targets. Immediately relevant is the effect of confusers. Adding the confusers reduces the number of points on target by 24% across all devices and targets.

Figures 6–15 present the target detection statistics for each target with and without confusers and for each lidar at each target. In each lettered lidar plot the horizontal line indicates the median and surrounding the box spans the 25th to the 75th percentile; the middle 50% of the total number of data points or IQR. Outliers are determined by finding the upper distance threshold, T_{upper}

$$T_{\text{upper}} = D_{\text{median}} + 2.5R_3, \tag{1}$$

and the lower distance threshold, T_{lower}

$$T_{\text{lower}} = D_{\text{median}} - 2.5R_2. \tag{2}$$

In Eqs. (1) and (2), R_2 and R_3 represent the second and third quartile ranges around the median. Outliers are those values that exceed these threshold values and are indicated as red "+" markers.

Figures are organized sequentially starting with 5 m targets and include both lanes. Examining the results for the 5 m target in Fig. 6, it can be seen that lidars A and H record no points on the 5 m target. These lidars both have an FoV in the range of 60 deg. In our test configuration, the 5 m target was situated at nearly 60 deg from the test axis connection between the reference center and the 200 m target; the target is likely just outside the FoV of both lidars these figures.

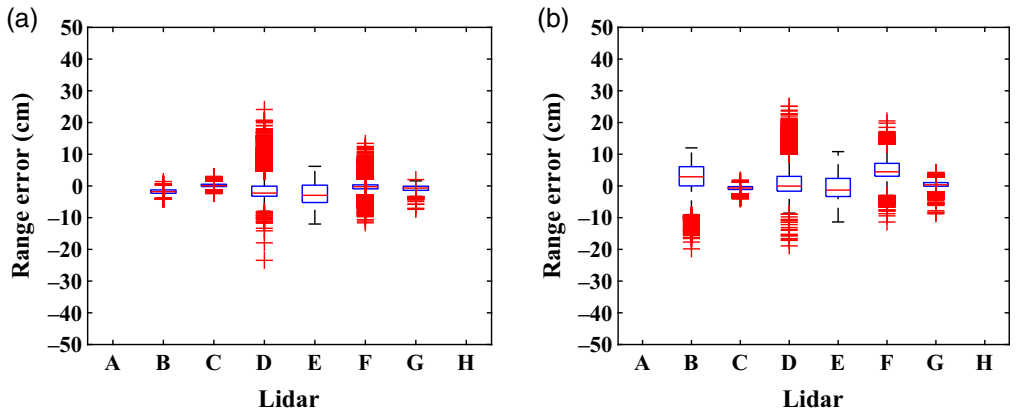

Fig. 6 Boxplots of range error by lidar for the 5 m target. (a) Without confusers and (b) with confusers.

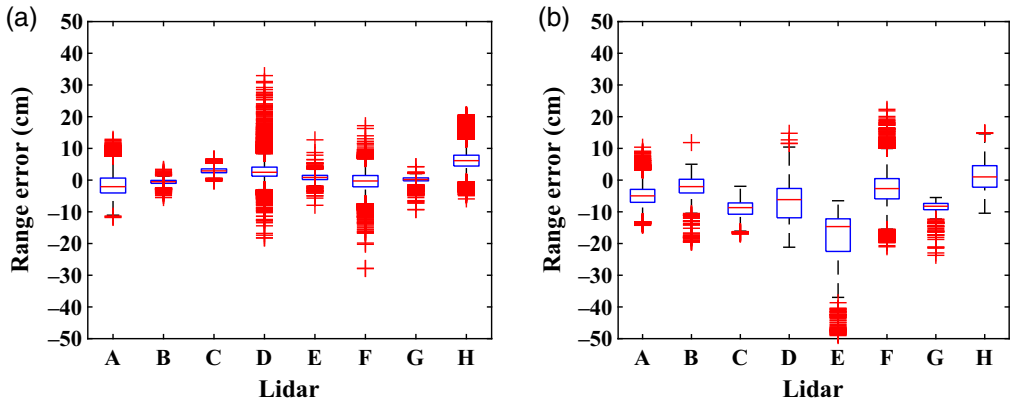

Fig. 7 Boxplots of range error by lidar for the 10 m target. (a) Results for the bare target without confusers and (b) with confusers.

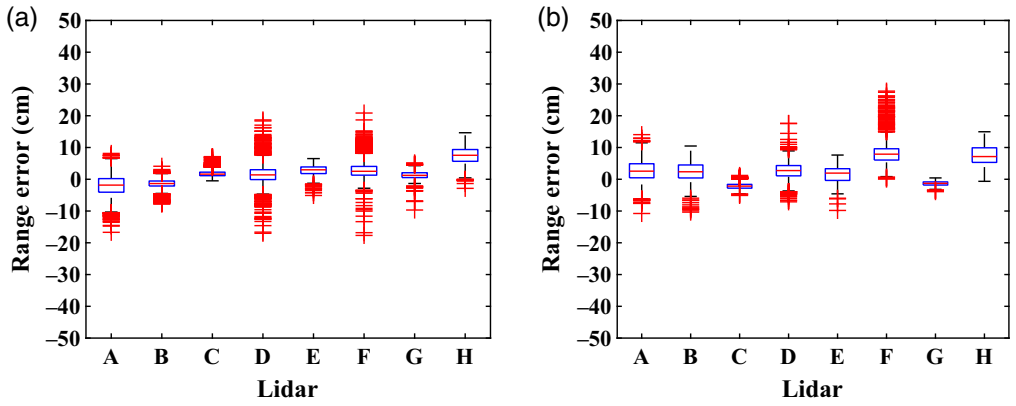

Fig. 8 Boxplots of range error by lidar for the 15 m target. (a) Results for the bare target without confusers and (b) with confusers.

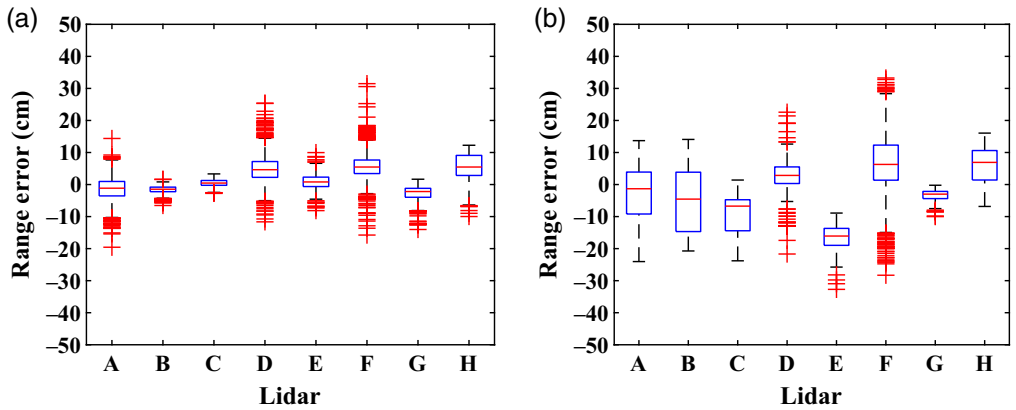

Fig. 9 Boxplots of range error by lidar for the 20 m target. (a) Results for the bare target without confusers and (b) with confusers.

While Tables 2 and 3 provide an overview of the performance of the test lidars as a group, Figs. 6–15 tell us much more about the variation in performance of each lidar compared the others. For example, in Fig. 6 we can see that in "Lane 1" lidars B and C have roughly the same performance when observing the 5 m target. When confusers are introduced the median reported range, IQR and number of outliers increase. While the IQR of lidar C increases from 0.7 to 0.9 cm lidar B's increases by 5 cm. A similar trend is observed in the performance of lidar B's scoring the other targets.

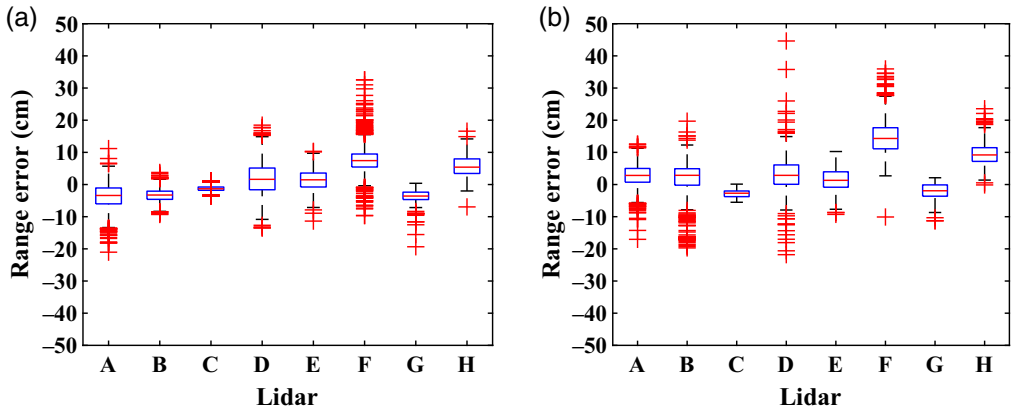

Fig. 10 Boxplots of range error by lidar for the 25 m target. (a) Results for the bare target without confusers and (b) with confusers.

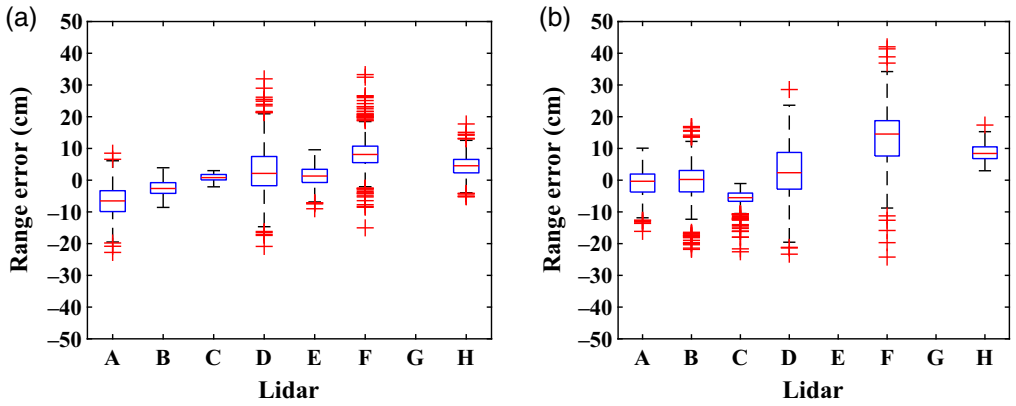

Fig. 11 Boxplots of range error by lidar for the 30 m target. (a) Results for the bare target without confusers and (b) with confusers.

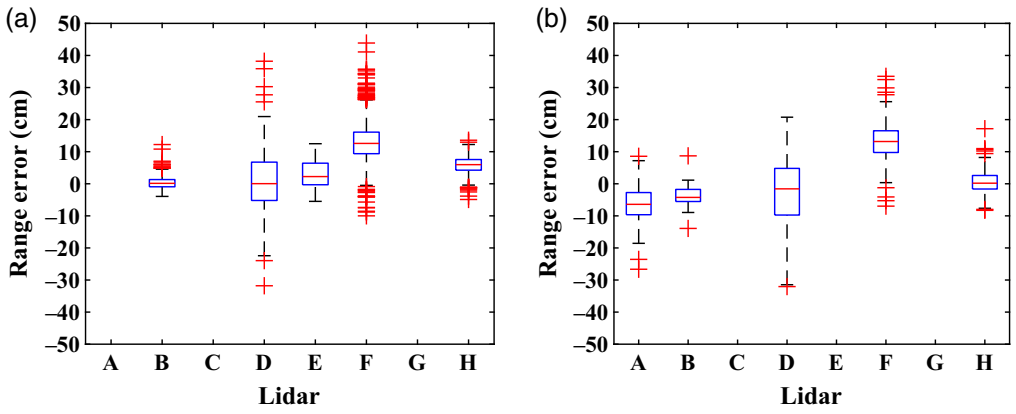

Fig. 12 Boxplots of range error by lidar for the 35 m target. (a) Results for the bare target without confusers and (b) with confusers.

In contrast, lidars C, D, E, and F show a decrease in the number of outliers in between Lanes 1 and 2. Lidar G, like lidar B, also shows and increases with the introduction of confusers.

The variation in some reported range values is also interesting to note: half of the values reported by lidar E differ from the median value by >5 cm whether confusers are present for the 5 m target or not. In fact, the difference between the minimum and maximum value reported

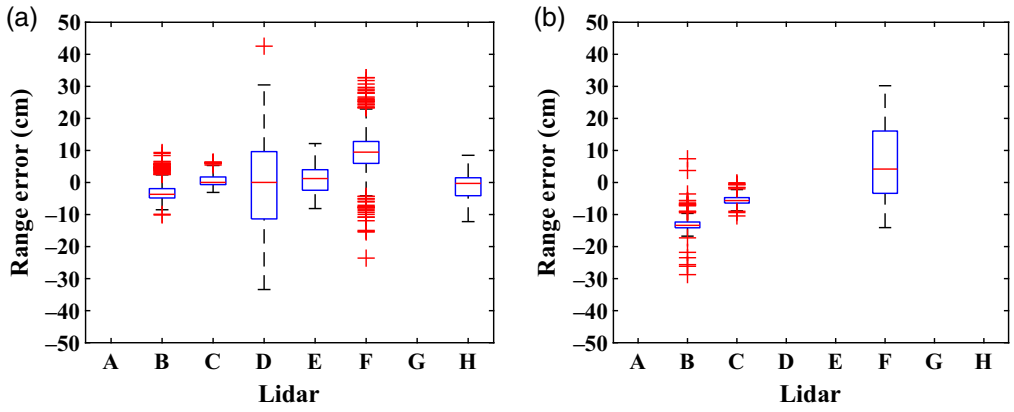

Fig. 13 Boxplots of range error by lidar for the 40 m target. (a) Results for the bare target without confusers and (b) with confusers.

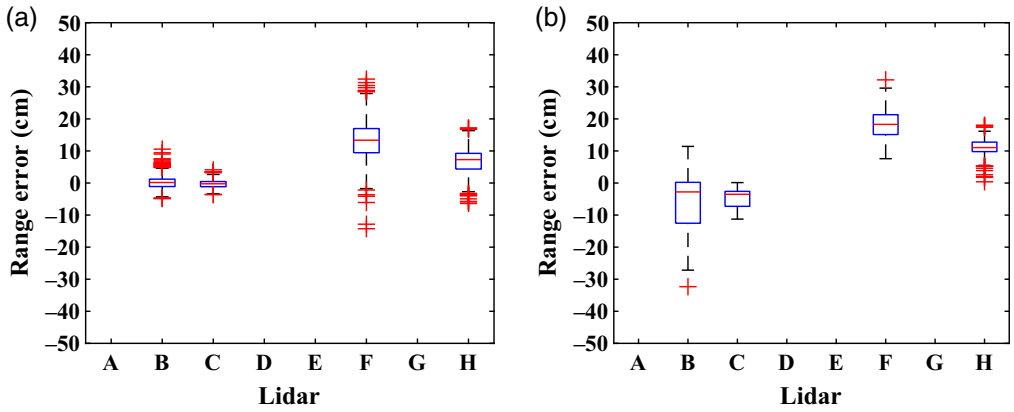

Fig. 14 Boxplots of range error by lidar for the 45 m target. (a) Results for the bare target without confusers and (b) with confusers.

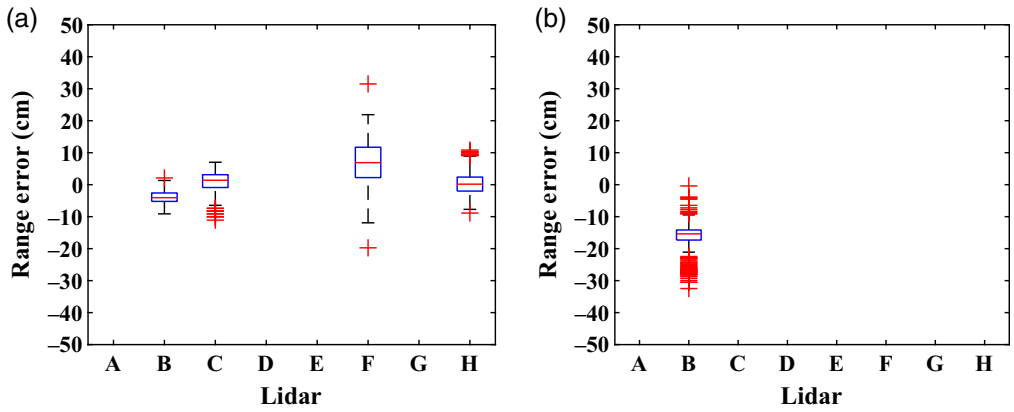

Fig. 15 Boxplots of range error by lidar for the 50 m target. (a) Results for the bare target without confusers and (b) with confusers.

for the 5 m target in Lane 1 was 18 cm over the 100 recorded scans. These results and others suggest a relatively fat-tailed probability distribution.

A similar box plot is shown in Fig. 7 here comparing the Lane 1 and Lane 2 results for the 10 m target. All units are represented in this plot. In Lane 1, we observe 50% of the reported points within about 2 cm of the median reported value for all DUTs. However, units D, F, and G report value of >10 cm from the median value 50% of the time, excluding outliers. For lidar A,

the observed deviation from the mean is 20 cm excluding outliers at the 50th percentile. Also of note is that 14% of the points reported by DUT E are considered outliers compared to around 8% to 9% for lidars G and H and compared with <3% for the other units. Here, again, we observe similar performance between units B and C and a reduction in the number of outliers reported after introducing confusers.

The trends observed in Figs. 6 and 7 continue at the 15 m target in Fig. 8. New here is the is an increase in the mean IQR by only 28% on average across all units compared from Lane 1 to Lane 2. By comparison at 5 and 10 m, the increase was 74% and 146%, respectively; the effect of the confuser is lessened. Conceivably, the confuser here could be far enough from the target as to not have as large an effect. This seems likely as the trend of continues again at 20 m in Fig. 9 and is observed generally in for the other range targets. Indeed, the IQR of DUT F increases from 4.2 to 10.9 cm for the 20 m target with the introduction of confusers. For the same device, a similar increase is observed at the 30 m target; 5.2 to 12 cm. While the data suggests that confusers can severely reduce range precision we cannot rule out issues with the test set up as the strongest confirming results are on the left side (10, 20, 30, and 40 m targets) of the test range. It is conceivable that ambient illumination or the location of the sun relative to the target and DUTs affected results.

Extrapolating general trends are possible only between the 10 and 35 m targets where most of the DUTs score every target. For example, we observe in Lane 1 the average IQR of all DUTs increases from 3.6 to 9.2 cm across 10, 15, and 25 m target data. The IQR of DUT B in particular increases from 1.4 to 18.5 cm. The performance of these units is as one might expect, IQR increases in general with range. A notable exception can be found in lidar G where the IQR does not vary considerably between the two test configurations and is similar with range also. However, data are only available out to 25 m for this unit for both lanes.

Continuing through Figs. 12–15, we see a common trend of decreasing range accuracy and precision. However, this appears to be mostly correlated with a drop in the number of points on target, caused by a decreasing occupied solid angle and also the presence of confusers when relevant. The presence of confusers decreases the number of points on target for all lidars and by more than half in some cases. Confirming results can be found in Table 4 in the appendix.

The boxplots in Figs. 16–23 present the same data as Figs. 6–15 but by range for each lidar. This presentation makes clear the difference in performance for each lidar with and without the presence of confusers. DUT A (Fig. 16) is somewhat typical of the test group. In Lane 1, the DUT tend to slightly underestimate range in this case by about 3 cm on average. In Lane 2, the IQR increases slightly from 5 to 6.3 cm and average estimated range error increases to −0.3 cm – still short of the target.

The effect of confusers is probably most obvious in the results for DUT B. In Fig. 17, performance is seen to be remarkably consistent in Lane 1 out to 70 m. The opposite story is told by the results for Lane 2 in the same figure. Because of this inconsistency, we cannot rule out processing artifacts of even a spurious event on the test range; a strong gust of wind for example.

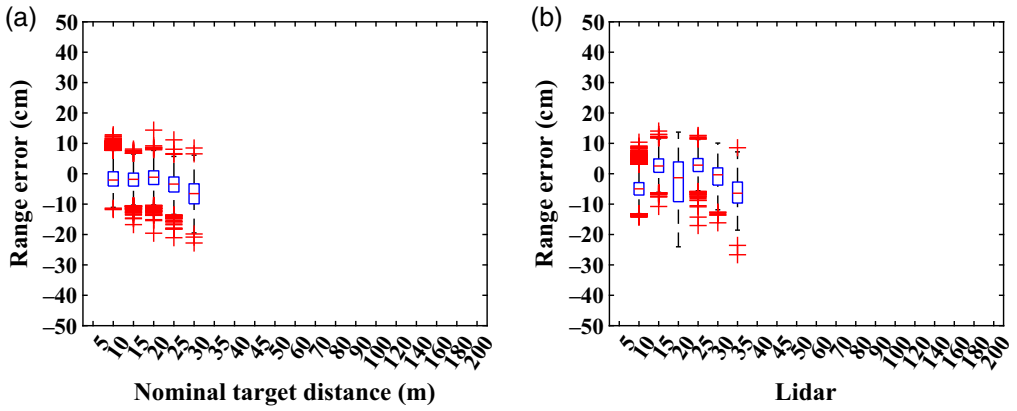

Fig. 16 Boxplots of range error by target for lidar A. (a) Results for the bare target without confusers and (b) with confusers.

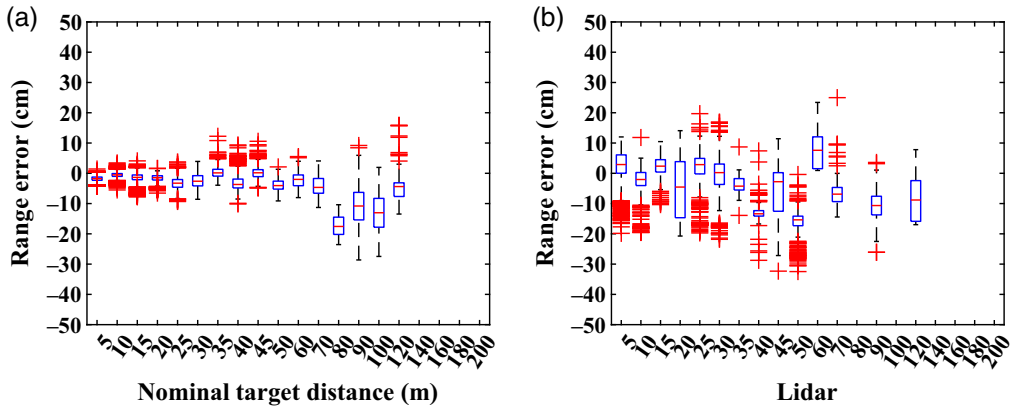

Fig. 17 Boxplots of range error by target for lidar B. (a) Results for the bare target without confusers and (b) with confusers.

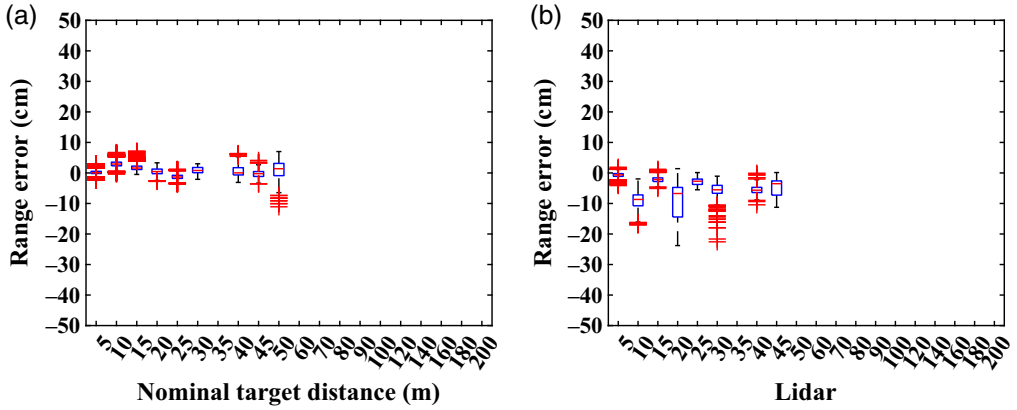

Fig. 18 Boxplots of range error by target for lidar C. (a) Results for the bare target without confusers and (b) with confusers.

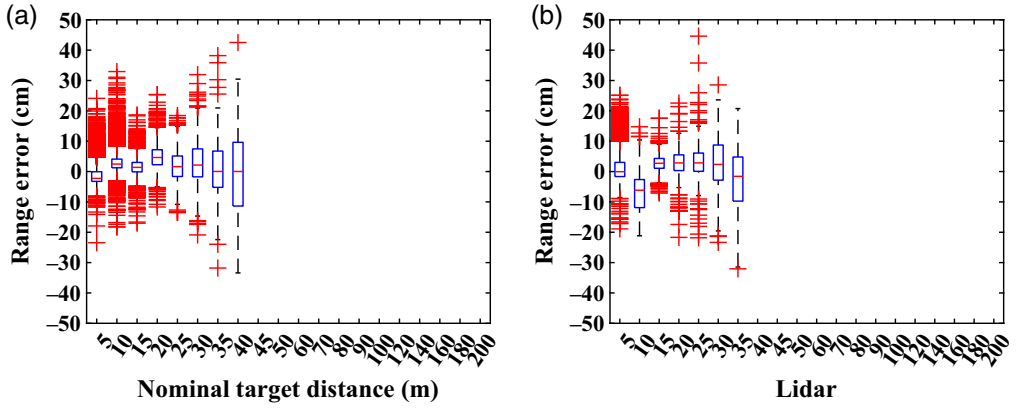

Fig. 19 Boxplots of range error by target for lidar D. (a) Results for the bare target without confusers and (b) with confusers.

Interesting over the remaining figures is the uniqueness of each lidar compared with the others. For example, if we exclude the 10 and 20 m targets DUT C consistently reports a similar range for these targets with high precision and is only slightly affected by the confusers. For DUT D, the IQR range increases sharply beyond 25 m in both tests. This drop is likely related to sharp drop in points on target (<400).

We suspect that other tests and metrics will allow a unique fingerprinting of lidar performance based on make or certain common design decisions. For example, if we were to assume

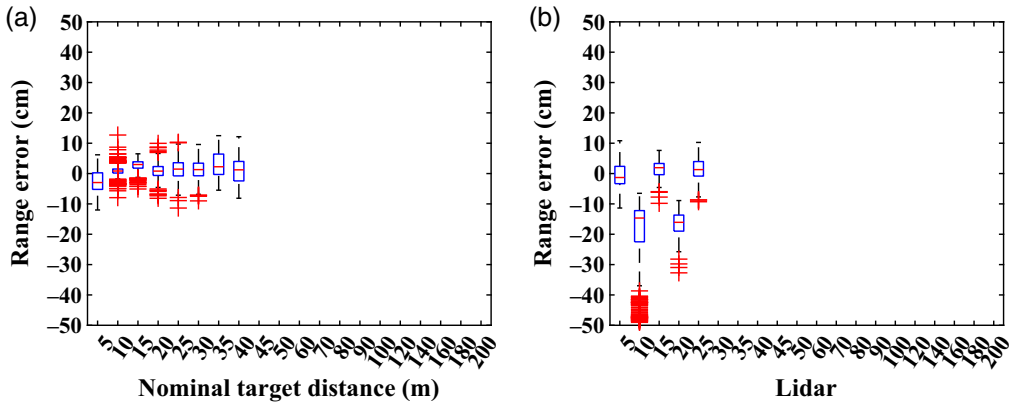

Fig. 20 Boxplots of range error by target for lidar E. (a) Results for the bare target without confusers and (b) with confusers.

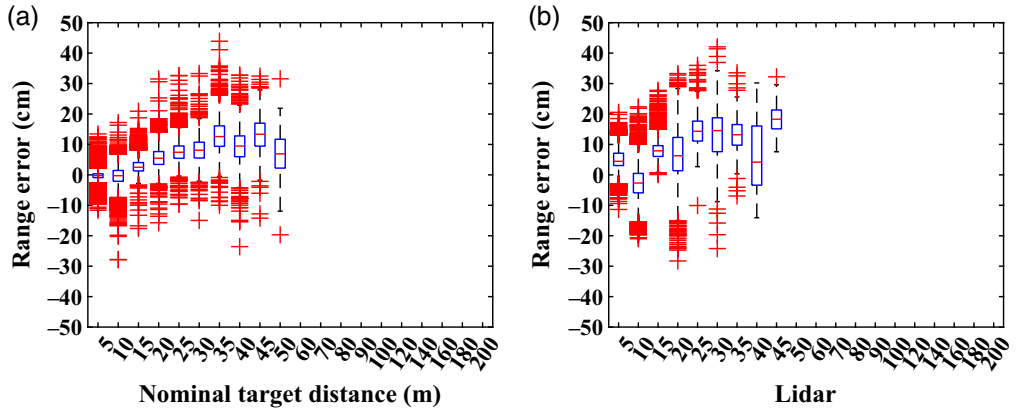

Fig. 21 Boxplots of range error by target for lidar F. (a) Results for the bare target without confusers and (b) with confusers.

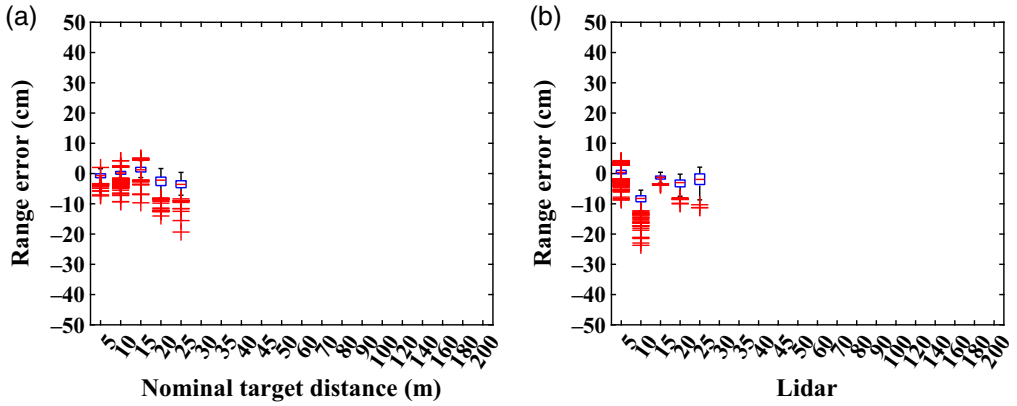

Fig. 22 Boxplots of range error by target for lidar G. (a) Results for the bare target without confusers and (b) with confusers.

that ambient lighting differences or sun-angle are driving the difference in the left/right sides of the test course than lidar E is may be said to be particularly affected by this difference; the same could be said of lidar F. In Fig. 21, the range error increases with distance with deviations from this trend and 10, 20, and 40 m.

The nature of the test setup also likely biased results against some of the units. Observe that in Tables 2 and 3 the total number points between lidars E and G are similar through E was able to

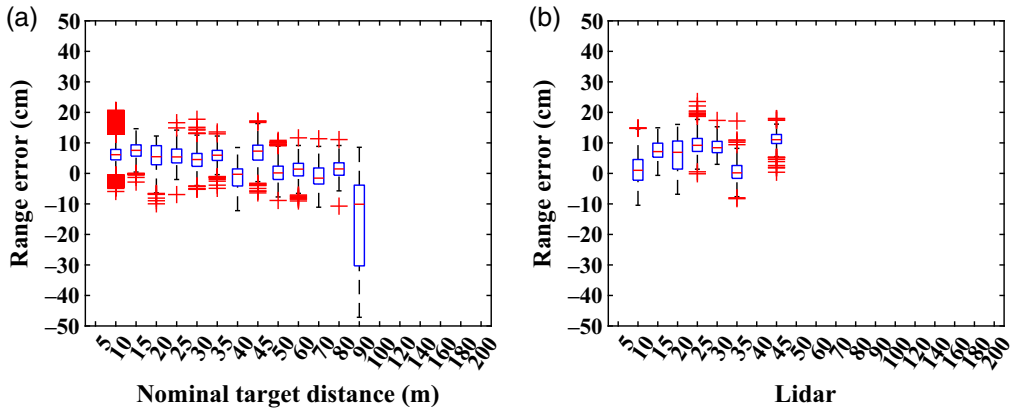

Fig. 23 Boxplots of range error by target for lidar H. (a) Results for the bare target without confusers and (b) with confusers.

detect targets out to 40 m. This is likely due to aspects of the test set-up favoring lidar E over lidar G in this instance. In this instance, vertical scanning of the lidars, larger, or longer targets may have been more equitable. Despite this, and commented upon previously, the performance of lidar G is notable. The IQR for lidar G increased by only 0.2 cm in the presence of confusers less than any other DUT. Lidar G also has most obvious skewness in distribution of detection samples.

One last unique performance characteristic can be observed in Fig. 23. Here, Lidar H was able to detect the 90 m target but reported the target nearly a meter closer compared to the reference and a with an IQR of 26.5 cm.

4 Conclusion

We have presented here the results of a first attempt at benchmarking eight automotive grade lidars. This effort is the first to use calibrated targets along with a reference and adjacent highly reflective confusers. Our purpose in this work is to motivate the development of test standards in this area and highlight variations in performance between lidars when stated specifications are similar. In this test, all the lidars operated near 900 nm, claimed either 100 m or 200 m maximum ranges, and range precisions of no worse than 3 cm as one standard deviation about the mean. Testing in this first year involved the first use of calibrated, Lambertian targets with 10% reflectivity in two test lanes with and without adjacent highly retro-reflective confuser targets.

In this first, early, initial examination of the test data we focused only on range accuracy and precision. This comparison was made possible via a survey-grade reference lidar. Across all devices tested we observed an average absolute range accuracy of 2.9 cm with respect to the reference across all targets. Average range precision was estimated at 3.6 cm. Introduction of the confusers in the second test decreased the number of points reported on target by 24% and increased range uncertainty by 34%. Additionally, the detection range, or range where 200 points were placed on target and averaged across all DUTs, decreased by half from 100 to 50 m. Only one DUT was able to detect targets beyond 90 m and the typical maximum range detection in our second test was 40 m. In addition, the results presented here indicate that, due to inherent design tradeoffs, the performance of each lidar is unique and can be characterized up to a point.

Generalizing across the tested devices, we can say that while the specifications listed by each vendor are representative, they do not adequately describe performance on their own; further justifying the need for standards. Advertised maximum range would appear to depend upon a very specific, and undocumented, set of circumstances coming together to detect a target. With respect to range precision, all the devices tested demonstrated precision similar to their stated specifications. However, it was common for the distribution of range estimates to be dominated by a concentration near the mean value and heavy tailed. Outliers were typically between 1% and 3% but sometimes as high as 14% for some units and targets. This finding may have

implications for object detection and tracking algorithms that assume detections will be normally distributed about the mean.

As we laid out in the introduction, the results presented here are from only the first year of proposed 3-year effort. However, there is still more to be done to improve our analysis of the year one data. Notably absent from this manuscript is an estimate of test uncertainty, repeatability, and reproducibility. Similarly, there are some inconsistencies in the test data that bear investigation. For example, there was tendency for detection on some targets on the left of the range to have a higher variability. It has been suggested that this may be due to background illumination or solar-angle; effects that have not been accounted for in the test setup.

To that end, proposed additions to the testing for years two and three are included in an appendix to this paper. These additions include more complex targets, dynamic targets, placing corner cubes, or identical lidars on the test range, and weather effects.

Years two and three also include plans to repeat the testing from previous years incorporating lessons learned. As we plan for year two those improvements for the Lanes 1 and 2 tests presented include maintaining the lidars in a power-on state prior to testing and ambient light monitoring/recording. Some changes to the test setup are also likely in order. Carefully orienting confusers adjacent to the targets and increasing the overall target height are likely to be considered. Other work includes improvements to the processing pipeline to accommodate a planned 30 lidars in year two. Finally, as we continue to develop these tests and standards a careful analysis of repeatability and error contributions is warranted.

5 Appendix A: Additional Tables

Table 4 contains aggregated statistics across all lidars, excluding the ground truth, for all targets. Table 5 provides detailed position and alignment information of each test target with respect to the ground truth lidar. The last column provides a double dot product misalignment loss factor.

Table 4 Average of all DUTs for each target.

	Points		Accuracy		IQR	
Target	Lane 1	Lane 2	Lane 1	Lane 2	Lane 1	Lane 2
5	24,573	25,656	−1.3	1.0	1.6	2.8
10	10,759	5317	1.3	−5.8	2.4	5.8
15	3323	2458	1.8	2.6	2.5	3.2
20	3096	1416	1.5	−2.0	3.6	9.2
25	1123	929	0.6	3.6	2.9	4.5
30	832	382	1.1	3.2	4.9	8.2
35	487	189	4.2	0.2	6.2	7.2
40	454	201	1.1	−4.9	6.5	4.6
45	433	258	5.1	5.8	4.1	6.6
50	304	195	1.1	−15.4	5.1	3.2
60	430		−0.3		3.8	
70	217		−3.1		5.1	
80	123		−8.0		4.9	
90	132		−10.5		17.8	
100	77		−13.0		9.5	

Table 5 Target alignment information.

Target number	Relative position (l/r)	Nominal distance (m)	Actual distance (m)	Nominal angle (deg)	Subtended angle (deg)	Subtended angle lower boundary (deg)	Subtended angle upper boundary (deg)	Target pitch (deg)	Target yaw (deg)	Target yaw relative to sensor (deg)	Potential misalignment loss
1	Right	5	5.021	−45.985	1.835	−46.921	−45.086	3.272	−46.377	−0.392	0.998
2	Left	10	10.031	19.438	0.936	18.955	19.892	3.175	10.142	−9.296	0.985
3	Right	15	16.062	−30.202	0.683	−30.536	−29.854	2.530	−33.033	−2.832	0.998
4	Left	20	20.307	15.253	0.474	15.024	15.498	6.314	0.669	−14.584	0.962
5	Right	25	25.788	−22.271	0.409	−22.487	−22.078	3.364	−17.777	4.493	0.995
6	Left	30	30.141	11.274	0.299	11.119	11.418	7.263	13.543	2.269	0.991
7	Right	35	35.283	−16.981	0.258	−17.112	−16.854	2.460	−17.062	−0.081	0.999
8	Left	40	40.394	9.537	0.243	9.428	9.671	5.068	9.207	−0.331	0.996
9	Right	45	44.700	−13.216	0.224	−13.328	−13.103	3.869	−10.980	2.236	0.997
10	Left	50	50.210	8.278	0.233	8.162	8.395	0.119	6.076	−2.203	0.999
11	Right	60	60.041	−10.735	0.181	−10.828	−10.647	5.745	−12.866	−2.131	0.994
12	Left	70	69.735	6.790	0.139	6.718	6.858	1.041	9.727	2.937	0.999
13	Right	80	79.208	−8.171	0.128	−8.234	−8.106	4.418	−9.618	−1.446	0.997
14	Left	90	89.619	5.119	0.126	5.058	5.185	7.739	5.233	0.114	0.991
15	Right	100	99.253	−5.768	0.111	−5.823	−5.711	9.097	−7.411	−1.643	0.987
16	Left	120	119.648	3.854	0.085	3.813	3.898	−0.801	6.012	2.158	0.999
17	Right	140	139.659	−2.363	0.067	−2.398	−2.331	7.365	−13.804	−11.442	0.972
18	Left	160	160.110	2.350	0.057	2.325	2.382	8.842	1.302	−1.049	0.988
19	Right	180	181.317	−1.382	0.050	−1.408	−1.358	10.854	−22.771	−21.389	0.914
20	Center	200	202.326	1.337	0.046	1.315	1.361	5.158	−0.366	−1.703	0.996

6 Appendix B: Year 2 Expected Approach

Major goals of year two testing include examining eye-safety and interference and is planned for Friday, April 28 and Saturday, April 29, before the SPIE DCS conference the following week at the same facility as the year one test. We tentatively expect 30 lidars designed for the automotive sector. Prior to on-site testing, each lidar will undergo a series of eye safety measurements conducted by Exciting Technology (ET) in their Dayton, Ohio, optical labs. The purpose of these eye safety tests is to determine the nominal ocular hazard distance (NOHD) for each lidar for both unaided and aided viewing. The general concept is to integrate the output of a high bandwidth InGaAs detector over a time period of 10 s for various distances from the lidar. The aperture size of the unaided optics will be ~7 mm, representative of a dark adapted eye, and 50 mm in diameter for the aided optics, representative of a 50 mm diameter binocular. Table 6 below summarizes the required eye safety measurement equipment.

Ideally, we would use a single detector for every lidar wavelength, but the range of wavelengths manufacturers used could be quite broad. We have initially selected an extended range InGaAs detector from LabSphere with responsivity from 800 to 2600 nm, peaking at 2200 nm. This particular detector as more spectral range than is required for this application, as the highest wavelength lidar, we anticipate testing is 1550 nm. Figure 24 contains all the information we currently have for the detector. Before the test, we would need the responsivity specifications for each lidar wavelength as well as the temporal bandwidth, as sub-ns rise/fall times (ideally DC− $\geq$ 2 GHz) are desired to capture peak power from pulsed lidars for more accurate integration.

Table 6 Required eye safety equipment.

Equipment	Description	Supplier
DUT	Device under test – the lidar itself	Various
PSU	Power supply unit – specific to each DUT	Various
Interface Cabling	Required cables for each DUT/PSU and interfacing with OS	Various
IDA-EXT-050-RTA-CX	Extended range InGaAs detector assembly. 800 to 2600 nm.	Labsphere
Frame Adapter	0.5″ frame adapter for InGaAs detector	Thorlabs/Edmund
Unaided Optics Tube	Custom lens tube to simulate 25 mm aperture	Thorlabs/Edmund
Aided Optics Tube	Custom lens tube to simulate aided eye 75 mm aperture	Thorlabs/Edmund
Oscilloscope	High bandwidth sampling oscilloscope	ET
Computer	Data storage and lidar control	ET

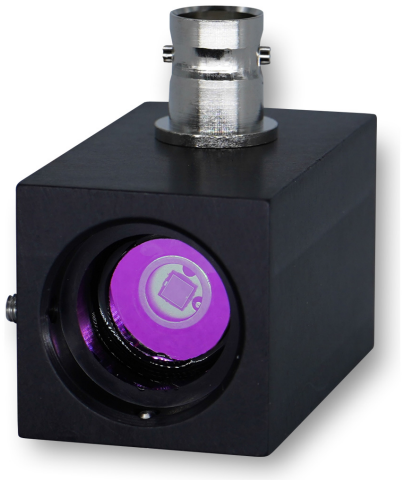

Fig. 24 Labsphere IDA-EXT-050-RTA-CX InGaAs detector and specifications.

For the on-site lidar field testing, we will implement two new lane configurations. The first test lane will combine the unconfused and confused lanes of the year one tests – leveraging the observation that confusers not close in proximity to targets had no effect on the targets detection. Additional elements such as cement barricades, simulated tire fragments and simulated negative obstacles (Positive obstacles are convex relative to the ground plane. Negative obstacles are concave and are more difficult to detect.) may also be included. The second test lane will examine the impact and susceptibility of the various DUTs to interference events.

While interference may occur organically during the course of driving or naturally due to certain environments, it may also occur intentionally and possibly maliciously. Another test will be primarily designed to identify naturally occurring interference effects on automotive lidars. Interference may manifest itself as false positives or ghost targets that do not physically exist at the detected location, false negatives or missing targets, or detected targets shifted in position. Each of these cases can cause potentially dangerous results.

In order for interference to occur, the victim DUT must receive an interference event within its range gate time and the DUT must be spatially aligned with the source in some way. For direct interference, the devices need to be spatially aligned so the two DUTs are facing each other with overlapping FoVs. For indirect interference, an interfering alignment can occur with the DUTs imaging the same object at the same time, and the victim lidar must interpret the received interference as a target. Depending on the receiver architecture, it may have some resistance to interference and the resistance can differ between architectures. Specifications regarding interference are not typically released by manufacturers.

Despite best efforts, individual lidars will not be phase locked to each other. Thus, the temporal alignment between lidars can be modeled at random, or quasi-periodic at best. The typical motion of lidars mounted on vehicles is expected to be moderately complex. While cars generally move in translational motion, rotational motion is also expected due to the vehicles' suspension during motion, pot holes, and natural curvature of the driving experience, among others. Furthermore, many commercial lidars scan patterns result in a dynamic but repeatable scanning mechanism to cover a scene. As a result, the FoVs of the lidars will point at each other and overlap during some instances.

Static interference testing will generally follow the procedure outlined in Popko.[17] Each lidar will individually be tested against each of the other DUTs as both the victim lidar and the interfering lidar. The existing test setup in year two will utilize lane one with the addition of a second pedestal for the interfering lidar to be placed. The height should correspond to a typical location on a vehicle. Future efforts will place each lidar on a stage that will permit a naturally occurring rotational motion (yaw, pitch, and roll).

During each test, a point cloud will be collected from the victim lidar while the interference lidar is turned off. After scanning the scene for a fixed amount of time, this point cloud will be considered the 'truth' for the victim lidar. Next, the experiment will be repeated but with both the victim and the interfering lidar on. Point clouds will be compared with any new or shifted targets (identified based on a tolerance value to be determined) and quantified. The number, location (space and angle), repetition, etc., of return points will be characterized in a confusion matrix. Primarily, the deviation from the "interference-free" case will be noted as lidars may have vastly different error and accuracy metrics.

The methods above will certainly allow us to evaluate interference from a limited amount of lidars. We will also examine using large corner cubes to simulate returns from identical, oncoming lidars, but this is the pathologically worst case scenario because the signal is fixed in range and is at exactly the same frequency and in phase. We will also examine using modulated corner cubes, such as NRL has used for two way communications, as this may simulate a more realistic interference scenario. If we can validate the test methodology of using large corner cubes we can easily simulate many more interfering lidars.

7 Appendix C: Year 3 Expected Approach

In year 3, we plan on replicating the tests of year 2, incorporating lessons learned, and additionally testing DUT's susceptibility to weather-related performance degradation.

We wish to develop test protocols and associated metrics to measure the performance of lidars under varying weather conditions, such as fog (the international definition of fog is visibility <1 km (35 db/km attenuation at 0.5 km, visibility ~92.3% transmission at 10 m) and mist is visibility between 1 and 2 km (8 db/km attenuation at 1.5 km, visibility ~98.2% transmission at 10 m)[18] and rain (rain intensity is defined by the US Geological Survey as light rain falling at <0.5 mm/h, moderate rain between 0.5 and 4.0 mm/h, heavy rain between 4.0 and 8.0 mm/h, and very heavy rain in excess of 8.0 mm/h[19]). We will implement repeatable test conditions that simulate weather events for the duration of the test and maintain uniformity across the testing platform and evaluate lidar performance by measuring the reflectivity of chosen targets under various test conditions at various fog and rain intensities. The lidars will be tested under at least two different fog levels and rain intensities, ideally moderate and heavy. Additionally, we propose to test lidars during both simulated rain events where the surfaces are wet and there is rain actively falling, and again after a rain event when only the surfaces are wet.

The weather tests could take place during the same test event or at a separate facility at a different location, but using an existing facility that supports weather testing with a wide range of control on the testing parameters may be the best option. Various state departments of motor vehicles (DMVs) have testing facilities used to examine vehicle safety under real-world weather and lighting conditions. For example, Virginia Tech's Smart Roads Program[20] has a 2.2 mile highway section with controlled lighting and weather systems that are capable of producing fog and rain of varying intensities and droplet sizes, as shown in Fig. 25. Arranging test times at such facilities is possible but would require additional travel and logistics.

The Naval Research Laboratory's Laboratory for Autonomous Systems Research (LASR) facility[21] in Washington, DC, is a smaller testing facility that could be used. The facility contains a $40' \times 60' \times 46'$ tropical high bay that simulates a south-eastern Asian rain forest as seen in Fig. 26. The temperature is held constant at 80°F with 80% humidity. The tropical high bay is capable of producing fog and rain with varying rates up to 15 mm/h. A catwalk at 15′ level along the perimeter allows access to mounting equipment with a separate observation room that provides dry space for electronics and computers.

If dedicated facilities are unavailable, limited weather conditions can be simulated nearly anywhere using commercially available hardware. To simulate fog, we propose to use a fog generation system such as a pulley drive mist pump[22] that we have utilized in laboratory testing. Sandia National Lab uses one such system,[23] shown in operation in Fig. 27 to replicate low visibility fog by driving water through a series of standard misting nozzles in buckets via hoses connected to a single pump. Two green laser beams are used to measure the transmission through the fog and verify its uniformity. In order to sustain such a fog, the humidity of the enclosed area

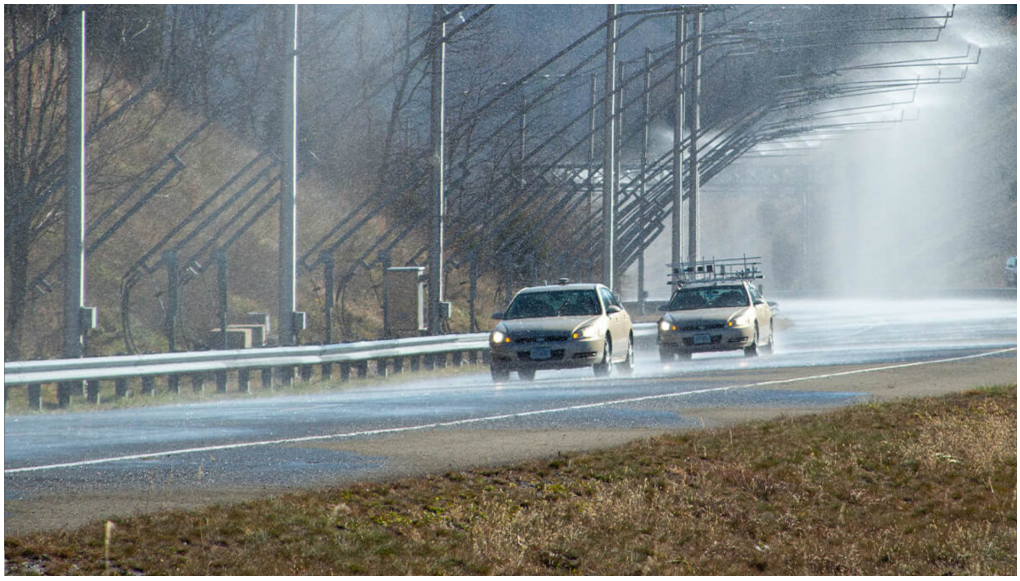

Fig. 25 The rain testing section of Virginia Tech Transportation Institute's highway section.[20]

Fig. 26 The LASR facility at Naval Research Laboratory features a tropical high bay capable of simulating rain and fog.[21]

Fig. 27 The fog facility at Sandia National Laboratory, using lasers to measure and calibrate transmission.[23]

has to be maintained at >80%. This requirement poses a challenge for outdoor testing, as the tents, we aim to use for such weather testing need to allow for almost complete closure. Transmission measurements at specified spatial intervals should be implemented to monitor the visibility conditions as a function of space and time (day-to-day) to ensure uniform testing conditions during the relevant test periods.

Not having prior experience with rain simulation, we plan to adapt simple rain machine techniques used in the film industry. Commonly available hardware can be used to build a network of PVC pipes with various sprinkler heads to achieve optimum coverage of the testing area as shown in Fig. 28. The flow rate of water through the system controls the amount of rain produced and the droplet size. Control over the flow rate is needed if we wish to evaluate the performance

Fig. 28 A simple rainfall simulator commonly implemented in the film industry (Tom Antos Film).

of lidars under different rain conditions. The rain drops distribution needs to be monitored as a function of time to ensure consistent conditions exist throughout the experiment. Droplet size can be measured by direct imaging (shadowgraphy) or a phase-Doppler anemometer, but the equipment are not at hand and need multiple units to monitor the extent of the test site. Setting up such a system and testing to ensure repeated and controlled performance is imperative for successful implementation. Prior experience testing autonomous lidars suggest that distances on the order of 10s of meters may be required before measurable performance degradation can be observed. A long-range test facility to simulate rain with good control and equipment for testing droplet size distributions at set intervals will be a challenge.

Acknowledgments

The authors of this paper wish to express their gratitude to Riegl USA for providing the VZ-400i TLS and supporting equipment, Labsphere Inc. for proving the calibrated reflectance targets, SPIE for supplying the traffic cones and road signs, BRIDG and NeoCity for providing the test location, and the following volunteers for helping with data collection efforts: Randy Arabie and Peter Hallett (SPIE), Ian Mattson and Matthew Spencer (Michigan Tech), Chris Durell and Savannah Labounty (Labsphere, Inc.), Gabe Hallett (University of Washington), and Cullen Bradley and Eddie Ruff (Exciting Technologies). Anna Hecht, from Exciting Technology, contributed to the section on eye safety and Chris Valenta (Georgia Tech Research Institute) and contributed the section on interference testing in Appendix B. Vasanthi Sivaprakasm, (U.S. Naval Research Laboratory), contributed to the section on weather in Appendix C. All authors declare that they have no conflicts of interest.

References

1. D. Carruth et al., "Predicting error propagation in autonomous ground vehicle subsystems," in *Proc. Ground Veh. Syst. Eng. and Technol. Symp.*, pp. 11–13 (2020).
2. C. Glennie and D. Lichti, "Temporal stability of the velodyne HDL-64E S2 scanner for high accuracy scanning applications," *Remote Sens.* **3**(3), 539–553 (2011).
3. C. Glennie and P. Hartzell, "Accuracy assessment and calibration of low-cost autonomous lidar sensors," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **XLIII-B1-2020**, 371–376 (2020).
4. M.-A. Mittet et al., "Experimental assessment of the Quanergy m8 lidar sensor," in *ISPRS 2016 Congr.* (2016).

5. M. Kutila et al., "Benchmarking automotive lidar performance in arctic conditions," in *IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, IEEE, pp. 1–8 (2020).
6. P. Rosenberger et al., "Benchmarking and functional decomposition of automotive lidar sensor models," in *IEEE Intell. Veh. Symp. (IV)*, IEEE, pp. 632–639 (2019).
7. P. Rosenberger et al., "Towards a generally accepted validation methodology for sensor models-challenges, metrics, and first results," in *Proc. 2019 Graz Symp. Virtual Vehicle*, Graz, Austria, pp. 1–13 (2019).
8. S. Cattini et al., "A procedure for the characterization and comparison of 3-D LiDAR systems," *IEEE Trans. Instrum. Meas.* **70**, 7002110 (2020).
9. J. Lambert et al., "Performance analysis of 10 models of 3D LiDARs for automated driving," *IEEE Access* **8**, 131699–131722 (2020).
10. J. Kim et al., "Performance of mobile LiDAR in real road driving conditions," *Sensors* **21**(22), 7461 (2021).
11. J. Schulte-Tigges et al., "Benchmarking of various LiDAR sensors for use in self-driving vehicles in real-world environments," *Sensors* **22**(19), 7146 (2022).
12. E57 Committee, "Standard test method for evaluating the relative-range measurement performance of 3D imaging systems in the medium range," Tech. Rep., ASTM International, West Conshohocken, Pennsylvania (2015).
13. E57 Committee, "Standard test method for evaluating the point-to-point distance measurement performance of spherical coordinate 3D imaging systems in the medium range," Tech. Rep., ASTM International, West Conshohocken, Pennsylvania (2017).
14. Federal Highway Administration, *Manual on Uniform Traffic Control Devices for Streets and Highways*, 2009 ed., US Department of Transportation (2009).
15. Riegl Laser Measurement Systems, GmbH, *Riegl VZ-400i Technical Manual*, rev. 2018-02-20 ed., Riegl Laser Measurement Systems, GmbH, Horn (2018).
16. Stanford Artificial Intelligence Laboratory, "Robotic operating system," https://www.ros.org (2018).
17. G. B. Popko, T. K. Gaylord, and C. R. Valenta, "Interference measurements between single-beam, mechanical scanning, time-of-flight LiDARs," *Opt. Eng.* **59**(5), 053106 (2020).
18. N. Islam Md and M. N. Al Safa Bhuiyan, "Effect of operating wavelengths and different weather conditions on performance of point-to-point free space optical link," *Int. J. Comput. Netw. Commun.* **8**(2), 63–75 (2016).
19. P. H. US Geological Survey USGS Water Science School, "Rainfall calculator, metric-How much water falls during a storm? USGS Water Science School," https://water.usgs.gov/edu/activity-howmuchrain-metric.html (2022).
20. Virginia Tech Transportation Institure, "Virginia Smart Roads | Virginia Tech Transportation Institute," https://www.vtti.vt.edu/facilities/virginia-smart-roads.html (2022).
21. U.S. Naval Research Laboratory, "LASR Facilities," https://www.nrl.navy.mil/lasr/facilities/ (2022).
22. Fogco Environmental, "Pulley Drive Mist Pump .5 GPM 1HP 115V 9 FLA 566 RPM | Fogco Environmental Systems," https://fogco.com/product/pulley-drive-mist-pump-5-gpm-5hp-115v-8-5-fla-566-rpm/ (2022).
23. Sandia National Laboratories: News Releases?: Testing sensors in fog to make future transportation safer," https://newsreleases.sandia.gov/fog_tests/ (2022).

Zach Jeffries is a PhD candidate in the Department of Electrical and Computer Engineering at Michigan Technological University and a member of the Robust Autonomous Systems Lab. He is a member of SPIE and also a Department of Defense SMART Scholar. As a SMART Scholar, he collaborates with the U.S. Army on lidar-based autonomous off-road navigation. He received his MS degree from Michigan Technological University and his BS degree from Loras College in 2020 and 2018, respectively.

Jeremy P. Bos is an associate professor in the Department of Electrical and Computer Engineering at Michigan Technological University. He is a senior member of SPIE, Optica, and IEEE as well as the author on over 100 scholarly works in the areas of atmospheric optics, scene

recovery, and autonomous vehicles. Before beginning his current position he was an NRC post-doctoral research associate for the US Air Force Research Lab. He received his PhD, MS, and BS degrees in 2012, 2003, and 2000, respectively. He has 10 years in the automotive and defense industries before returning to academia and is a licensed professional engineer in the state of Michigan.

Paul McManamon is president of Exciting Technology LLC, chief science officer at Nuview, and technical director of the Lidar and Optical Communications Institute, LOCI, at the University of Dayton. He retired from being chief scientist for the Air Force Research Lab, AFRL, Sensors Directorate in 2008. He chaired the US National Academy of Sciences Study "Laser Radar: Progress and Opportunities in Active Electro-Optical Sensing" (2014). He was the main LiDAR expert witness for Uber in the lawsuit vs Google/Waymo. He was co-chair of the US NAS study "Optics and Photonics, Essential Technologies for Our Nation" (2012) study, which recommended a National Photonics Initiative, NPI. He was also vice chair of the 2010 NAS study called "Seeing Photons: Progress and Limits of Visible and Infrared Sensor Arrays." He has written two books on Lidar. He is a Fellow of SPIE, IEEE Optica, AFRL, DEPs, MSS, and AIAA. He received the WRG Baker award from the IEEE in 1998 for the best paper in ANY refereed IEEE journal or publication (>20,000 papers). He was president of SPIE in 2006. He was on the SPIE board of directors for 7 years and on the SPIE Executive Committee from 2003 through 2007. Prior to being chief scientist for AFRL Sensors Directorate he was senior scientist for EO/IR sensors, and acting chief scientist for the Avionics Directorate for >2.5 years. In 2006 he received the Meritorious Presidential Rank Award. He was the co-recipient of the SPIE Presidents' Award in 2013.

Biographies of the other authors are not available.