Data Readiness Levels for Automated Driving

Ian Marsh[1], Victor Stenmark[2] Yuri Poledna[3],
Heikki Hyyti[4], Martin Sanfridson[1], Eren Aksoy[5]

***Abstract*— Automated vehicles need to drive at least as well as humans using sensors, data, and machine intelligence. Therefore, this paper investigates data quality and its impact on upstream machine learning. Clearly, poor data increases the risk of hazardous situations and requires additional cycles in the drive-gather-process-assess process, increasing costs and probably delays. A complicating factor for data quality is adverse weather, prevalent in Nordic conditions.**

This work introduces a novel measure for quantifying the data quality. Akin to Technical Readiness Levels, Data Readiness Levels have been developed within this work to quantify data quality. The idea is a simple overall measure of the dataset, ranging from 1-9. As well as a framework, DRLs normalise values and metrics from various assessment tools.

To evaluate the concept, three autonomous driving datasets were used, augmented with additional weather conditions not present in some scenes. Two of the datasets are from open-road driving scenes, and one from a test track. Images were further distorted with rain and fog, and point clouds with fog. All source code is available and, one data set is, with the other two part of a larger release.

We show that a simple, single metric can be derived from 46 image and 2 point cloud quality metrics tested across nearly 5 TB of data. Through considerable statistical rankings, two image quality metrics performed well, at some computational cost, whilst state of the art point cloud metrics were inconclusive across our data.

Keywords: **Data Quality, Automated Driving, Image Quality Assessment and Point Cloud Quality Assessment.**

I. Introduction

This paper systematizes degradations in multimodal data sources for Automated Driving (AD). Quality degradations, or impairments, were assessed from sensor installation at the hardware level to object detection at the machine learning level. The focus of this paper, however, is on the image and point cloud quality using open source metrics. Our motivation in this work is to ensure the data gathered during the data collection phase is complete and of sufficient quality. Simply, iterating the data collection phase is costly, and when incomplete potentially dangerous when the AD vehicle is in production and use.

[1]Research Institutes of Sweden (RISE) AB, Sweden. firstname.lastname@ri.se.

[2]The Royal Institute of Technology, Sweden. vstenm@kth.se.

[3]CARISSMA Institute of Automated Driving, Technische Hochschule Ingolstadt, Ingolstadt, Germany. yuri.poledna@carissma.eu.

[4] The Finnish Geospatial Research Institute (FGI), National Land Survey of Finland. heikki.hyyti@nls.fi.

[5]Lund University, Sweden. eren.aksoy@cs.lth.se.

Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them. Project grant no. 101069576.

Fig. 1: The data gathering vehicle used evaluating DRLs in this paper. The driven journey is summarised in Section V-A. Photo credit: The Finnish Geospatial Research Institute FGI.

To attribute the data quality, or Data Readiness Level (DRL) to a dataset, is in principle relatively straightforward. One has to conceive a framework in which to assess the data, identify the processing stages, obtain sufficient and relevant data, select the conditions in which an AD vehicle may find itself and finally process and summarise the data.

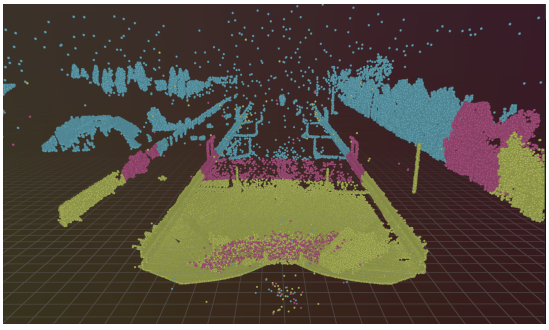

Fig. 2: A LiDAR point cloud from our test track in Ingolstadt, Germany. The point cloud is augmented by noise to evaluate point cloud quality metrics. Fog has been artificially added to the point cloud. Attenuation, one parameter in fog generation, is shown by colours in the image.

II. Related Work

From 2024, Liu et al. postulate that AD dataset surveys are not sufficiently extensive [1]. Therefore, the authors present an exhaustive study of 200+ datasets encompassing sensor modalities, varying sizes, tasks, and contextual conditions. Of

relevance is that they introduce a metric for evaluation. From 2020, Vargas et al. look at autonomous vehicle sensors and their vulnerability to weather conditions [2]. From 2021, Yurtsever et al. present challenges, high-level system architectures, emerging methodologies, and core functions, including localisation, mapping, perception, planning, and human-machine interfaces [3]. The nuScenes publication, from 2020, contains a comprehensive survey of AD datasets [4] as well as their dataset. Bijelic et al. from 2020 examine adverse weather conditions and present a novel multimodal dataset acquired in more than 100 km of driving in Northern Europe [5]. This dataset is the first large multimodal dataset encompassing adverse weather, with 100K labels for LiDAR, camera, RADAR, and gated near-infrared sensors. The Boreas dataset, from 2023, with 350 km driving data, was collected over a repeated route over one year, with seasonal variations and adverse weather conditions [6]. A list of AD datasets with weather considered is available [7].

Model-based NR-PCQA techniques extract colour and geometry information directly from 3D point clouds without projecting them onto a 2D plane [8]. Liu et al [9] proposed ResSCNN, a model-based NR-PCQA method that uses sparse convolutional layers. These layers are similar to regular convolutional layers except that they only operate on the occupied points of the 3D point cloud, which reduces the computational load [10]. ResSCNN uses sparse convolutional layers for feature extraction, followed by pooling and concatenation of the feature vectors, and finally predicts the scores using two fully connected layers. It achieved state-of-the-art performance on NR-PCQA benchmarks, outperforming both FR-PCQA and other NR-PCQA methods [9].

III. AD, Sensing and Data Quality

RGB imaging and quality: All AD vehicles have at least one RGB camera, often several, where the combination of cameras with a predetermined focal distance enables stereo vision. In this work, level 6 of the DRL is derived from image quality assessment (IQA), specifically **no reference** image quality. That meant the quality was assessed without a reference, or in machine learning terms, the ground truth. Our chosen metrics used three image attributes:

1) Focus: Blurred-Sharp
2) Exposure: Light-Dark
3) Contrast: Low-High

Thermal imaging: Thermal cameras detect radiation, which is essentially heat that surrounding objects emit. Given the heat, thermal cameras detect living objects better than non-living, animals, pedestrians, vulnerable road users as well as warm engines / exhaust. Moving, rather than parked (cooled), vehicles may be better detected. *LiDAR:* The LiDAR available was a Velodyne VLS-128, released in 2017. LiDAR is at Level 6, and point clouds at Level 7. The reason is processing can be done on the point cloud, with operations such as de-noising or filtering. The test track had a 4D Radar, whilst the open road driving did not use Radars. *Radar:* Radar typically improves on RGB cameras and LiDARs in adverse weather conditions. Snow and drizzle attenuate the wavelengths of light used in LiDAR, and deposits on the laser surface have an additional effect. In AD, the Radar and processing hardware produce point clouds for analysis, called '4D' Radar from range, azimuth, elevation, and velocity.

IV. Data Readiness Levels

A. The concept

Figure 3 shows nine levels with respect to data quality within automated driving. The focus of this work, levels 6-7, a complete description of Readiness Levels was introduced in [11], our contribution is available from GitHub[1].

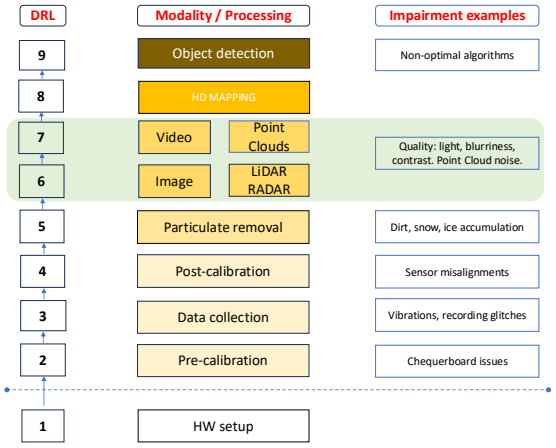

Fig. 3: *Leftmost:* Data Readiness Levels. During training, data flows from DRL 2, calibration, upwards to object detection, DRL 9. *Central*: indicate discrete steps aligned with the DRLs. *Rightmost*: example impairments. The shaded green section indicates this papers' focus.

Levels 6 and 7 are the subject of this paper. Each modality is handled separately and then combined into a single scalar value, as formalised in Algorithm 1.

B. DRL implementation

Levels 6-7 in the DRL structure hold the main contribution within this paper. Initially 46 (IQA) and 10 point-cloud (NR-PCQA) quality tools were selected. Throughout the results section, these were narrowed down to five image and two point cloud metrics. Each metric is then used from images and point clouds from the two datasets described.

The scores are combined, downsampled and processed into a time series, and then averaged, algorithm 1. The $\alpha, \beta, \gamma, \delta$ parameters are assigned equal weights. λ, or decay value was set to 0.9. This value tracks score fluctuations, without over- or undershooting the score tracking. The sampling rate r was set to the lowest refresh rate, which is 10 Hz, the LiDAR scan rate.

[1] https://github.com/roadview-project/No_reference_image_and_point_cloud_quality

Algorithm 1 DRL-derived metrics from four modalities $D_{[1,4]}$ downsampled, weighted, time-averaged and output.

Input: Data streams $D_1(t), D_2(t), D_3(t), D_4(t)$
Parameters: Weights $\alpha, \beta, \gamma, \delta$; decay factor λ; downsample rate r
for $i = 1$ to 4 **do**
    Downsample $D_i(t)$ with rate r: $D_i^r(t) \leftarrow \text{Downsample}(D_i(t), r)$
    Initialize $EWMA_i \leftarrow D_i^r(1)$
    for $t = 2$ to N **do**
        $EWMA_i \leftarrow \lambda \cdot D_i^r(t) + (1 - \lambda) \cdot EWMA_i$
Combine weighted values:

$$DRL \leftarrow \alpha \cdot EWMA_1 + \beta \cdot EWMA_2 + \gamma \cdot EWMA_3 + \delta \cdot EWMA_4$$

Output: Combined value DRL

V. Real world drives + augmentations

Before using our testing sites, the DRL Image Quality Assessment (IQA) using scenes augmented by *constructed* weather conditions was calibrated, an example is shown in Figure 4. Therefore, weather impairments can be generated, rain and foggy artifacts on real images[2], which needs large amounts of data and for some metrics, considerable CPU/GPU computations, especially for those using ML-trained models of IQA and PC-IQA quality.

q

A. FGI: Finnish Geospatial Research Institute dataset, FI.

The dataset is divided into rural and urban drives in Finland in the winter. The data was recorded on 13th December 2023. Since the recording vehicle, see Figure 1, was driven on public roads in Europe, both faces and number plates were anonymised through blurring. The dataset, for both drives, was 2 x 7 GB for each 49 minute drive. Release forthcoming.

B. REHEARSE: Carissma Institute, Ingolstadt Uni, DE.

The REHEARSE dataset is derived from 2 test sites [12], in this work only the German part from Carissma was used. A test site means known-calibrated distances and elevations, digital mapping, on-site management, chequerboard calibration and weather generation. *Known conditions and distances assist DRL calibration.* The datasets were created to observe the three weather conditions rain, clear, and fog. The rain was measured and calibrated using litres/m^2, drop shape+amount, and direct weather measurements. This compares with our rain generation for image augmentation. Four sensor modalities were used: RGB and thermal imaging, LiDAR, and RADAR. The dataset is 68 min. long totalling 320 GB on disk as street-type scenes. Available from REHEARSE.

VI. Weather impairments

Despite having 5 TB of data, controlled 'noise' was needed. Noise means rain and fog for images, and fog for point clouds. That is where the point cloud data is read from a LiDAR or 4G Radar. As well as the controllability in datasets, the REHEARSE test-track dataset has the possibility to generate real rain, drizzle, but not really fog, at least that remains in an outside environment. Over 3M frames and 50K point clouds without augmentation were rated, in addition to the samples 'impaired' with weather conditions.

[2]Data, code & visualisation from Noise models, metrics, and computation.

A. Image augmentation

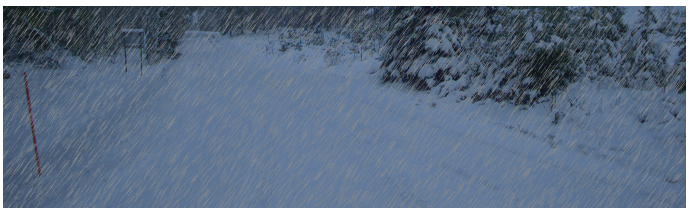

Fig. 4: Rain generation using the Albumentations package on the FGI dataset (subsection V-A) below. To calibrate the DRL Level 6, the impairments were controlled or varied, sometimes called 'noise modelling'. This particular image has been rain-impaired with a level of 20, on a scale from 0-100.

B. Point cloud augmentation

We used a method proposed by Hahner et al. referred to as `LiDAR fog sim` to synthetically augments point clouds [13]. It is a tool that in a physically accurate manner 'adds' to a point cloud [13]. It is done by tracing LiDAR light rays laser back & forth [14], [15]. An example was shown in Figure 2, details of the fog simulation for DRLs is in [16].

VII. Statistical evaluation

We use the same method to evaluate the results of the image and point cloud methods. No-reference quality assessments are assessed by comparing predicted scores with human evaluations for the same images or point clouds. Subjective scores in are given on a five-point scale [8], [17]. *The images* are either synthetically or authentically distorted. Synthetically distorted images are high-quality samples that have been artificially degraded by applying effects such as blur or compression. Authentically distorted images are those captured in real-world conditions, where distortions naturally arise due to poor lighting, motion blur, or lens effects. For *point clouds*, evaluation relies on synthetically distorted data, common distortions being Gaussian noise, colour noise, and compression effects.

NR-QA methods are typically assessed using the following four metrics: Pearson, Spearman, Kendall correlation coefficients and the Root Mean Square Error [8], [18], [19]. Pearson correlation and RMSE assesses the accuracy of the predicted scores by measuring their deviation from corresponding subjective scores. In contrast, the Spearman and Kendall coefficients evaluate how well the predicted scores preserve the rank order of the subjective evaluations.

Kendall, like Spearman uses ranks as opposed to absolute scores. It is computed by comparing all possible pairs of observations and classifying them as concordant, discordant, or tied. For two variables $X = \{X_1, ..., X_n\}$ and $Y = \{Y_1, ..., Y_n\}$, all pairs were considered, i.e. $\{(i, j) : i < j \leqslant n\}$. A pair is

concordant, discordant, or tied depending on:

$$\begin{cases} \text{Concordant} & \text{if } (X_i - X_j)(Y_i - Y_j) > 0 \\ \text{Discordant} & \text{if } (X_i - X_j)(Y_i - Y_j) < 0 \\ \text{Tie} & \text{if } (X_i = X_j) \oplus (Y_i = Y_j) \end{cases}$$

With these definitions, Kendall's correlation coefficient is:

$$\text{Kendall} = \frac{C - D}{\sqrt{(C + D + T_X) * (C + D + T_Y)}} \quad (1)$$

where C is the number of concordant pairs, D is the number of discordant pairs, T_X is the number of ties in X, and T_Y is the number of ties in Y. Worked examples are in [16].

VIII. Results

A. Image quality metrics

There are many IQA metrics. Only open-sourced ones were selected, with stars and pull requests on GitHub, and preferably accompanied by a peer-reviewed paper.

We used 46 metrics using a series of weather-impaired tests. Rain is in blue, and fog in orange. Note the behaviour of the metrics as the impairments increase. The next step to narrow down number of metrics is to correlate again against each other with respect to two conditions, namely rain and fog. This was done by three methods:

1) Finding metrics with similiar correlation values, under rainy and foggy conditions[3], figure 5.
2) Selecting metrics close in a metric space, figure 6.
3) Differing technology basis, see table I.

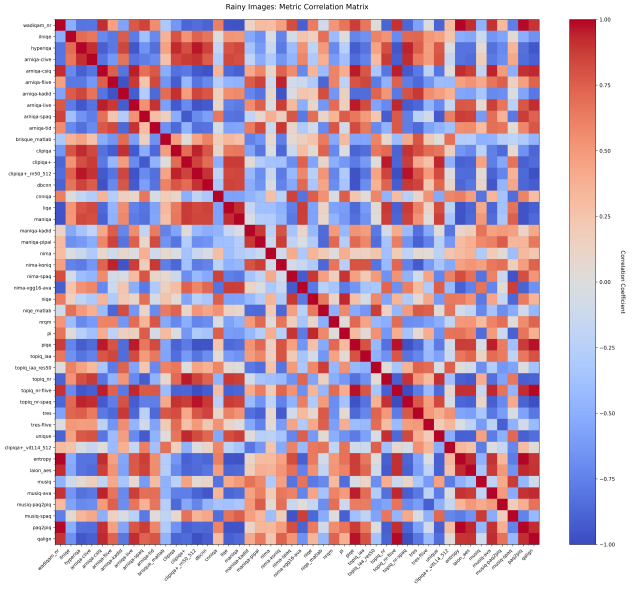

Fig. 5: A 46 image metric symmetric correlation heatmap for *rainy* conditions. Similar colours signify similar metric responses, i.e. highly positive correlated metric values.

[3]For space reasons, only rainy data is shown.

The second is finding metrics that are close in a metric space. Metrics that are close in a multidimensional space can be considered similar across all test data, and therefore are candidates for being removed as redundant. Since a compact quality assessment is needed, or a minimum set of metrics, data reduction is desirable. Note the scales in Figure 6 are not linear in the high dimensional space but appear linear when projected onto a 2D Euclidean space. Stress in MDS is a goodness-of-fit measure that indicates how well the low-dimensional representation preserves the original distances between points. For this 2D plot, the stress level is 14%, which is a fair fit, given some small distortion.

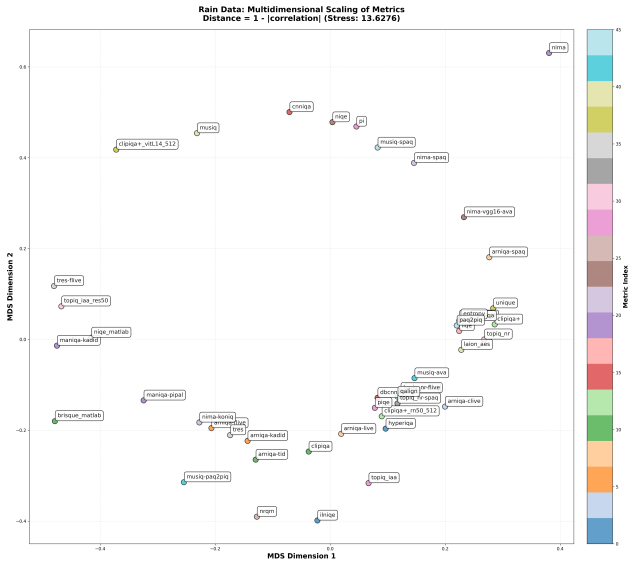

Fig. 6: Image Quality Assessment for 46 metrics. Point proximity indicates similarity in a Multi Dimensional Scale space. Stress indicates the goodness of fit on a 2D plot.

No.	Modality	Method	Technical Basis
1	Image	IL-NIQE	NSS-based
2	Image	DBCNN	CNN-based
3	Image	TOPIQ	Attention-based
4	Image	QualiCLIP	CLIP-based
5	Image	Q-Align	LMM-based
1	Point cloud	MM-PCQA	Projection and model-based
2	Point cloud	MS-PCQE	Projection-based

TABLE I: Five selected image quality metrics and two point cloud metrics. Note the different technical techniques used.

	Correlation	
Method	**Spearman**	**Kendall**
IL-NIQE	0.94 ± 0.1	0.86 ± 0.2
DBCNN	0.13 ± 0.9	0.16 ± 0.8
TOPIQ	-0.55 ± 0.7	-0.44 ± 0.6
QualiCLIP	-0.06 ± 0.7	-0.07 ± 0.7
Q-Align	0.99 ± 0.0	0.97 ± 0.0

TABLE II: Spearman & Kendall correlations for NR-IQA across distortion types & locations, see [16].

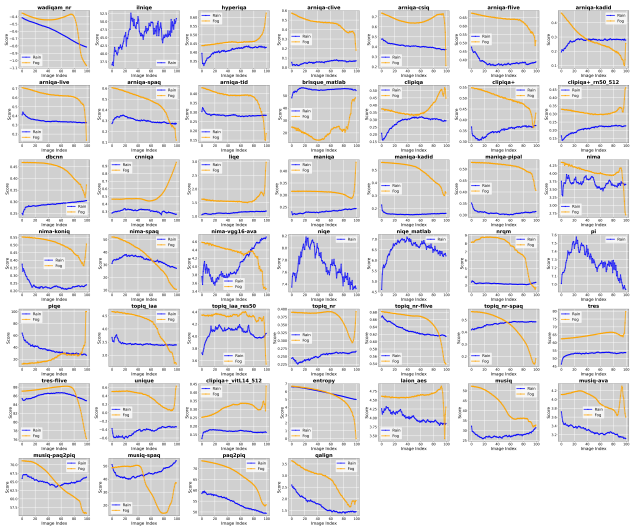

Fig. 7: Rain and fog compared for 46 metrics (Q-align best).

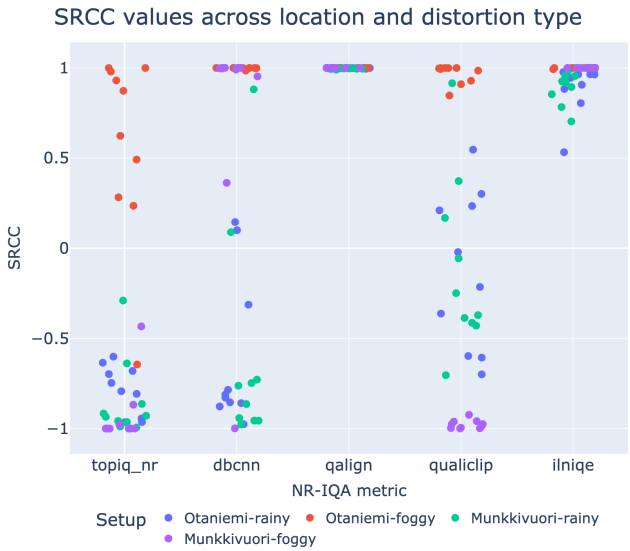

Fig. 8: Spearman Rank Correlation Coefficients (SRCC) for 5 NR-IQA methods. From drives in Finland and added rain (Figure 4) and fog situations using the vehicle in Figure 1.

B. RGB and thermal camera DRLs

The DRLs were derived from Q-Align over a series of images, scaled to 0-100, downsampled to the LiDAR refresh rate (10Hz), exponentially weighted, 0.1 recent values and 0.9 historical values and rounded to a scalar DRL 7 (images) and DRL 4-5 for the 3 thermal cameras using Algorithm 1. We do not give a DRL for the point cloud in this case.

C. DRL evaluations: RGB, thermal and point clouds

Figure 11 shows 4 views of our test track in Ingolstadt, Germany, with the DRLs marked. Higher values of the DRL indicated the image quality is adequate. Note, only a sole image is shown, however the rating is derived from long sequences. At longer physical distances, the area of focus is generally

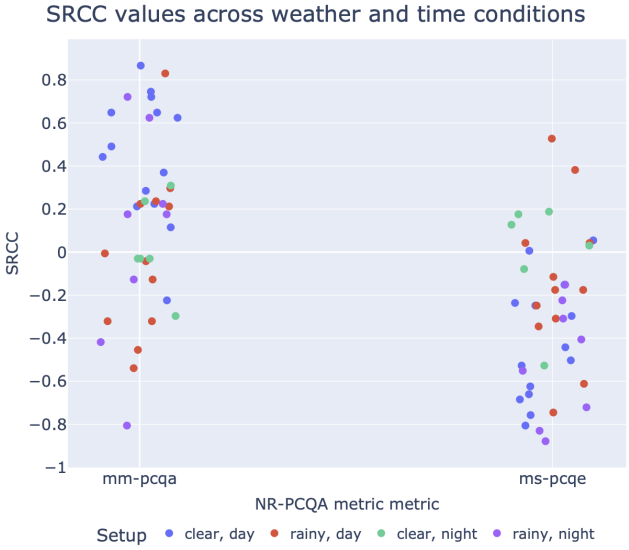

Fig. 9: Distribution of Spearman Rank Correlation Coefficient (SRCC) for 2 NR-PCQA methods from scenes from the test track in Germany. A spread of the results shows no clear correlation of human and derived values.

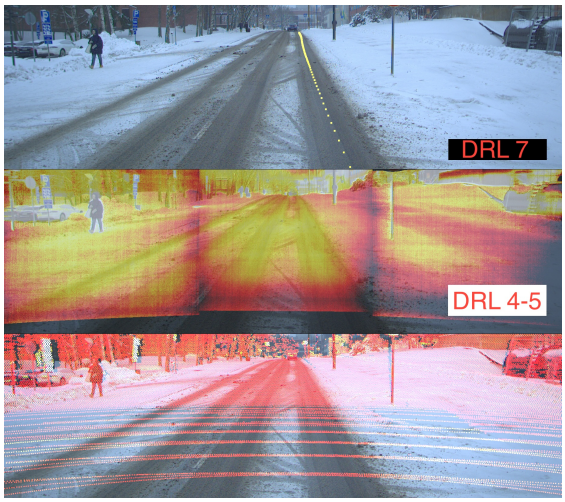

Fig. 10: Two data readiness levels superimposed on the Finish Geospatial Institutes (FGI) dataset. *Top:* an RGB road image with an estimate of the grip in dotted yellow. *Center:* three thermal camera images stitched together. *Bottom:* LiDAR range, and reflectance superposed on the top image. This is an example of setups that can influence data quality. It is not a judgement of thermal cameras quality, or this vehicle, or setup, which was rectified later.

better, and the centre image portion is focus. Generally, the sensing distance and range impacts the quality of the images. The Ingolstadt controlled environment can be instrumented to reflect the environment, but conditions can still differ: snow, sun, light levels etc. Investigating scenes requires a tool to start, stop, fast forward and rewind specific portions. Identifying the

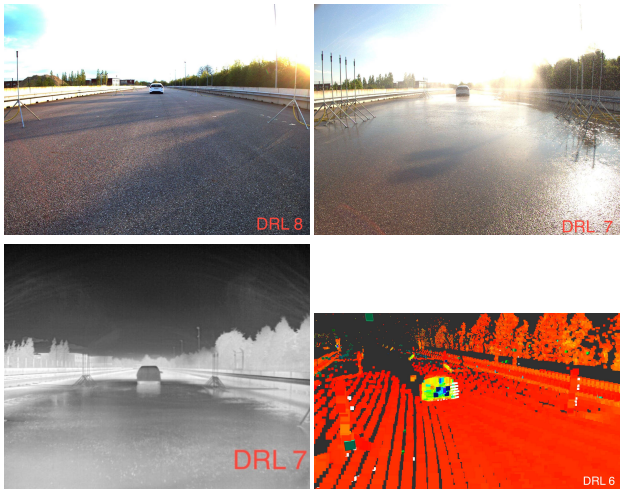

Fig. 11: Four images from the test track. *Top left:* DRL 8. A vehicle at 28m, an overexposed sky. *Top right:* DRL 7. RGB image of a vehicle at 28 meters sprayed in rain. *Bottom left:* DRL 7. A thermal image in black and white, a little less susceptible to some effects. Single thermal camera means no edge/fringe effects but offers less field of view. *Bottom right:* DRL 6. LiDAR point cloud. Reflection effects from the LiDAR are visible, the safety barriers and the vehicle.

same points in images and point cloud monitors how DRLs perform. An AD sequence with DRL example timelines is in this video sequence rerun [20].

IX. Discussion

Algorithm 9 shows an equal weight for each modality. It makes sense to make these, at least, function of time, i.e. for RGB camera(s) $\alpha \rightarrow \alpha(t)$. Sensor quality could be weighted, or accounting for a malfunction during an experiment. The computational cost of Q-Align is unsurprising, it does perform best in our chosen no-reference tools. A deeper investigation into performance versus should be done with tools such as weights and biases. No reference point cloud quality assessment methods were inconclusive for our data. AD point clouds differ in image scenes, objects, as well as sparsity in the training dataset. Models for the NR-PCQA are not perturbed by weather. AD point clouds contain noise, in general, dirt, dust, droplets. Sanitised point clouds for ML are not realistic enough for AD, and appropriate data, closer to real situations.

X. Conclusions

Autonomous driving *relies* on data. Poor data has been shown not only to jeopardize safety, but may also require additional time, effort, and cost (re)gathering, processing plus merging in newly gathered data. It is known high quality data will result in good machine learning performance. Correlation analysis of DRLs with object detection are ongoing. After exhaustive searches and implementations, five image quality packages and two point-cloud packages were selected. Regarding image quality, brightness, sharpness, and contrast are sufficient to assess the quality, however the techniques to obtain a sound metric as the quality deteriorates (linearly).

References

[1] Mingyu Liu et al., "A survey on autonomous driving datasets: Data statistic, annotation, and outlook," 2024.
[2] Vargas, Jorge et al., "An Overview of Autonomous Vehicles Sensors and Their Vulnerability to Weather Conditions," *Sensors*, vol. 21, no. 16, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/16/5397
[3] Ekim Yurtsever et al., "A Survey of Autonomous Driving: Common Practices and Emerging Technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.2983149
[4] Holger Caesar et al., "nuScenes: A Multimodal Dataset for Autonomous Driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11 618–11 628.
[5] Mario Bijelic et. al, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 11 679–11 689, 2020.
[6] Burnett et al., "Boreas: A multi-season autonomous driving dataset," *The International Journal of Robotics Research*, vol. 42, no. 1-2, pp. 33–42, 2023.
[7] I. Marsh, "Autonomous driving datasets and weather," Nov. 2025. [Online]. Available: https://ianmarsh.org/autonomous-driving-datasets
[8] S. Porcu, C. Marche, and A. Floris, "No-reference objective quality metrics for 3d point clouds: A review," *Sensors (Basel, Switzerland)*, vol. 24, no. 22, p. 7383, 2024.
[9] Y. Liu, Q. Yang, Y. Xu, and L. Yang, "Point cloud quality assessment: Dataset construction and learning-based no-reference metric," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–26, 2023.
[10] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky, "Sparse convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 806–814.
[11] N. D. Lawrence, "Data readiness levels," 2017. [Online]. Available: https://arxiv.org/abs/1705.02245
[12] Poledna et. al, "adveRse wEatHEr datAset for sensoRy noiSe modEls," 2024. [Online]. Available: https://s3.ice.ri.se/roadview-WP3-Warwick/T3.2%20-%20Create%20Dataset/rehearse/download.html
[13] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, "Fog simulation on real lidar point clouds for 3d object detection in adverse weather," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 283–15 292.
[14] L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 994–20 006.
[15] Y. Dong, C. Kang, J. Zhang, Z. Zhu, Y. Wang, X. Yang, H. Su, X. Wei, and J. Zhu, "Benchmarking robustness of 3d object detection to common corruptions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1022–1032.
[16] V. Stenmark, "Evaluating NR-IQA & NR-PCQA Methods on Weather-Distorted Data in Autonomous Driving," Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS), 2025.
[17] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Information Fusion*, vol. 91, pp. 230–260, 2023.
[18] M. et. al, "No-reference image quality assessment: Past, present, and future," *Expert Systems*, vol. 42, no. 3, p. e13842, 2025.
[19] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020.
[20] "Rerun SDK," 2025. [Online]. Available: https://www.rerun.io