Degree Project in Technology

Second cycle, 30 credits

Evaluating NR-IQA and NR-PCQA Methods on Weather-Distorted Data in Autonomous Driving

VICTOR STENMARK

Evaluating NR-IQA and NR-PCQA Methods on Weather-Distorted Data in Autonomous Driving

VICTOR STENMARK

Master's Programme, Computer Science, 120 credits
Date: July 13, 2025

Supervisors: Domova Veronika, Ian Marsh
Examiner: Haibo Li
School of Electrical Engineering and Computer Science
Host company: RISE
Swedish title: Evaluering av NR-IQA och NR-PCQA metoder på data med väderinducerat brus inom autonoma fordon

Abstract

In autonomous driving, balancing the quality of training data is essential for developing safe and reliable self-driving models. Exclusively using clear-weather data can reduce performance in adverse conditions, while too much bad-weather data may impair performance in clear conditions. A necessary prerequisite for striking this balance is to accurately assess data quality. This thesis investigates methods for evaluating the quality of images and point clouds without relying on reference data, known as no-reference image quality assessment (NR-IQA) and no-reference point cloud quality assessment (NR-PCQA), respectively. Five NR-IQA methods (IL-NIQE, TOPIQ, DBCNN, QualiCLIP, and Q-Align) and two NR-PCQA methods (MM-PCQA and MS-PCQE) were evaluated. The image data were sourced from the FGI dataset, which consists of two drives in winter conditions, and the point clouds were obtained from the REHEARSE dataset, comprising sensor data captured under controlled weather conditions. The images and point clouds used in the thesis were synthetically distorted with artificial fog and rain. The NR-IQA and NR-PCQA methods were evaluated based on their ability to accurately rank different versions of the same image or point cloud by distortion level. Neither of the two NR-PCQA methods demonstrated reliable performance in ranking the distorted point clouds. While MM-PCQA outperformed MS-PCQE, it did not produce accurate rankings. Of the evaluated NR-IQA methods, Q-Align and IL-NIQE achieved the best performance, significantly outperforming the other methods. These findings suggest that large multimodal models and natural scene statistics are promising approaches for assessing the quality of weather-distorted images. The results also suggest that NR-PCQA methods are not yet mature enough to reliably evaluate point cloud quality in the autonomous driving domain.

Keywords

Autonomous Vehicles, No-Reference Image Quality Assessment, No-Reference Point Cloud Quality Assessment, Weather-distorted Images, Weather-distorted Point Clouds, Data Quality

Sammanfattning

Vid träningen av självkörande fordon är det väsentligt att balansera kvaliteten på träningsdatan för att utveckla säkra och tillförlitliga modeller. Att enbart träna på data insamlad under goda väderförhållanden begränsar modellernas prestanda i dåliga väderförhållanden. Samtidigt kan träning på en för stor mängd data från dåliga väderförhållanden försämra prestandan under goda förhållanden. För att uppnå en balans i träningsdatan krävs tillförlitliga metoder för att avgöra datakvalitet. Det här arbetet undersöker metoder för att evaluera kvaliteten av bilder (NR-IQA) och punktmoln (NR-PCQA) utan användning av referensdata. Fem NR-IQA metoder (IL-NIQE, TOPIQ, DBCNN, QualiCLIP, and Q-Align) och två NR-PCQA metoder (MM-PCQA and MS-PCQE). Bilddatan hämtades från FGI, en datamängd som omfattar två körningar i vinterlandskap. Punktmolndatan hämtades från REHEARSE, en datamängd med punktmoln från en körbana. Bilderna och punktmolnen förvrängdes med syntetiskt regn och dimma. NR-IQA och NR-PCQA metoderna utvärderades utifrån deras förmåga att korrekt rangordna olika versioner av samma bild eller punktmoln efter förvrängningsgrad. Varken MM-PCQA eller MS-PCQE lyckades tillförlitligt rangordna punktmolnen. Även om MM-PCQA visade bättre prestanda än MS-PCQE var dess rangordningar mycket inkonsekventa. Bland de evaluerade NR-IQA metoderna uppvisade Q-Align och IL-NIQE den bästa prestandan. Resultaten tyder på att stora multimodala modeller och natural scene statistics är lovande metoder för att avgöra kvaliteten av väderförvrängda bilder. Dessutom tyder resultaten på att NR-PCQA metoder inte är tillräckligt utvecklade för att tillförlitligt utvärdera kvaliteten på punktmoln inom autonom körning.

Nyckelord

Autonoma fordon, No-Reference Image Quality Assessment, No-Reference Point Cloud Quality Assessment, Väderstörda bilder, Väderstörda punktmoln, Datakvalitet

Acknowledgments

I would like to sincerely thank my supervisor at RISE, Ian Marsh, for his valuable advice, support, and encouragement. I am also grateful to my academic supervisor Veronika Domova, for her guidance and feedback throughout the project.

Stockholm, July 2025
Victor Stenmark

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Autonomous vehicles (AVs) are becoming increasingly prevalent as more companies invest in self-driving technology. In 2024-2025, 45% of car manufacturers are either investing or planning to invest in autonomous vehicle technology [1]. Currently, the market for autonomous vehicles is valued at $122 billion, with projected revenues of $400 billion in 2035 [1].

The growing development of AVs is enabled by rapid advancements in deep learning and other technologies that make sophisticated autonomous systems possible [2]. The core of these systems lies in two key components: a sensor suite that collects data about the vehicle's surroundings and motion, and an autonomous driving system (ADS) that interprets this data to control the vehicle. Common sensors used in the sensor suites of AVs include cameras, LiDARs, and RADARs. Cameras capture visual information of the environment, which is typically used for tasks such as object detection and path planning. LiDARs and RADARs generate point clouds of the surrounding environment, enabling accurate estimation of distances to nearby objects.

A key challenge in autonomous driving is striking the right balance in the quality of the training data used to develop the ADS. Training exclusively on high-quality data captured in clear-weather conditions may limit the ADS's ability to generalize to environments with fog, rain, or snow [3, 4, 5]. However, relying too much on low-quality data, such as weather-distorted images and degraded point clouds, may degrade model performance in clear conditions [6]. To ensure robust ADS performance in production, it is therefore important to maintain a balance in the quality of data used during training, which in turn requires reliable methods for assessing data quality.

This thesis aims to improve the current understanding of evaluating the quality of both image and LiDAR data, with a particular focus on weather-

related distortions typical of Nordic conditions. Specifically, it compares methods for assessing the quality of images and LiDAR point clouds degraded by such conditions. The thesis is carried out at RISE in collaboration with ROADVIEW [7], an EU-funded project focused on developing perception and decision-making systems for automated vehicles in harsh weather conditions. ROADVIEW is interested in this project because it collects extensive image and point cloud data and needs to assess its quality to ensure the data is suitable for training autonomous driving systems. Additionally, the project interests autonomous driving researchers, who rely on accurate data quality assessment to develop autonomous vehicles with robust performance in adverse weather.

1.1 Research questions

- **RQ1:** What methods are suitable for assessing the quality of images captured by vehicle-mounted cameras under Nordic weather conditions?
- **RQ2:** What methods are suitable for assessing the quality of point clouds generated by vehicle-mounted LiDARs under Nordic weather conditions?

1.2 Purpose

The overarching purpose of this thesis is to enhance the robustness and reliability of autonomous vehicles. It aims to identify effective methods for image and point cloud quality assessment that can provide guidance for selecting appropriate training datasets for self-driving models.

1.3 Ethics and sustainability

Autonomous vehicles raise the ethical issue of algorithmic bias. AVs rely extensively on training data, and their performance is closely tied to the data on which they have been trained. As mentioned earlier, if the ADS has been trained solely on high-quality data captured in clear weather conditions, it is likely to perform poorly in situations with adverse weather, where data quality is typically worse. Given that autonomous vehicles are safety-critical systems, the ADS performing poorly in certain situations can have severe and even fatal consequences. This thesis aims to enable autonomous driving researchers to

accurately assess their data and identify a suitable mix of high- and low-quality data, ultimately leading to more robust autonomous systems.

There are multiple sustainability benefits that AVs might bring to society. AVs can optimize acceleration and braking patterns and reduce the time spent idling in intersections. Furthermore, shared fleets of AVs may reduce the need for car ownership, which in turn could decrease the number of cars on the road. Given the vast amount of resources necessary to manufacture vehicles, this could substantially reduce the resource use within the transportation sector.

1.4 Structure of the thesis

Chapter 2 outlines the context for data quality in the realm of AVs, weather-related distortions, and provides background on the fields of NR-IQA and NR-PCQA. It also discusses the related work to this thesis. Chapter 3 describes the selected methodology, chosen dataset, distortion techniques, and selection of image and point cloud evaluation methods. Chapter 4 presents the results of the evaluation of image and point cloud quality assessment methods. Chapter 5 discusses these results, and Chapter 6 summarizes the key findings and outlines possible directions for future work.

Chapter 2

Background

This chapter begins with an overview of data quality and the research fields of image quality assessment and point cloud quality assessment. It concludes with a review of related work relevant to this thesis.

2.1 Data quality in autonomous driving

Autonomous vehicles use a suite of sensors to navigate safely on the road, with LiDARs and RGB cameras being among the most widely utilized [8]. A LiDAR operates by emitting one or several laser beams and then processing the light signals that are reflected off the surrounding surfaces. [9]. In most LiDAR systems used in autonomous vehicles, distances to points are estimated using time-of-flight (ToF) measurements. This means that the distances to surrounding objects are calculated based on the time it takes for the laser pulse to travel to the surface and then return to the LiDAR. In addition to the measured distance, the LiDAR system records the corresponding azimuth and vertical angles, written as ϕ and θ, respectively. These angles, together with the distance measurement, are used to compute the 3D Cartesian coordinates of each point [9], creating a point cloud representation of the surrounding environment. Furthermore, the system typically records the intensity of each reflected laser pulse, which represents the strength of the echoed signal and is affected by factors such as the type of surface, range, and weather conditions [10]. This means that each point in the LiDAR output is represented in the form (x, y, z, i), where (x, y, z) denotes the Cartesian coordinate and i represents the return intensity. Cameras are also commonly used as sensors for autonomous vehicles, as they can capture images with rich visual details of the environment. This makes them particularly well-suited for tasks such as

object detection and classification, including the identification of road signs, pedestrians, and surrounding vehicles [11].

Several weather-related factors influence the quality of the collected images and point clouds. Rain can significantly impair camera performance as individual raindrops may stick to the lens, resulting in occlusions or blockages [12]. Object detection is also worsened in rainy conditions due to reduced brightness, obscured features, and rain streaks that cause visual noise [13]. Additionally, fog can cause uniform visual distortion, making it difficult to discern details and objects in the scene. Foggy conditions also create problems for LiDARs as the LiDAR beam is attenuated and backscattered [14]. Attenuation occurs when fog droplets absorb and scatter portions of the LiDAR's laser energy, leading to a reduction in signal strength. Backscattering occurs when part of the laser signal is reflected by fog particles, resulting in false returns and noise in the point cloud. LiDAR typically performs better in rainy conditions, as attenuation and scattering are negligible in low to moderate rainfall [12]. However, heavy and irregular rain causes lumps of fog, which leads to similar problems with attenuation and backscattering.

2.2 Synthetic data in autonomous driving

The use of synthetically distorted data in training autonomous driving systems has become common to improve their robustness in adverse weather. Several AV datasets consist exclusively of clear-weather data [15], which poses a problem because models trained on them tend to perform poorly in poor weather conditions [5]. One approach to addressing this problem is to collect new data in poor weather. However, collecting AV training data is both expensive and time-consuming. A more practical alternative used by researchers is to synthetically degrade clear-weather data to simulate adverse weather conditions.

Synthetic weather distortion of images takes multiple forms. One method of simulating rain involves overlaying an image with lines that mimic rain streaks. This method, used in the image augmentation tools `Albumentations` [16] and `imgaug` [17], allows the user to specify the amount of rain, as well as the size, shape, and slant of the raindrops. Both tools also support darkening the image, which is typically performed as rainy scenes tend to appear darker than non-rainy ones. `Albumentations` and `imgaug` also support fog distortion, which is implemented as large, semi-transparent circles overlaid on the image to simulate fog or haze. The opacity and blending factor of the fog can be adjusted, allowing the intensity of the fog effect to be

controlled. The primary advantage of these approaches is the ability to control all the parameters of rain and fog generation, which enables deterministic and reproducible results. However, these augmentation methods also have limitations. Neither `Albumentations` nor `imgaug` considers depth and assumes that raindrops are uniform in shape and size. More sophisticated methods for rain and fog generation take this into account. Halder et al. [18] developed a physics-based rain and fog generation tool that utilizes depth maps, camera calibration, and estimations of scene lighting to generate more realistic weather effects. The downside of this approach, however, is that it requires a per-pixel depth map, which is not always available.

Synthetic weather distortions have also been proposed for point clouds. Hahner et al. [19] proposed a computationally efficient fog simulation method that captures both attenuation and backscattering effects from fog particles. Their method operates on clear-weather point clouds by modifying the range and intensity of each point based on a validated optical model of a LiDAR. Hahner et al. used their fog distortion tool to fine-tune 3D object detection models on synthetically fog-distorted point clouds and evaluated the models on the Seeing Through Fog dataset, a dataset with authentic fog-distorted LiDAR data. The results demonstrated that training 3D object detection models on synthetically fog-distorted point clouds improves performance on authentically fog-distorted point clouds.

Using synthetic weather distortions to enhance robustness has also been used for images. Sakaridis et al. [20] developed a synthetic fog distortion method using depth information to enhance scene understanding, an umbrella term encompassing tasks such as object detection and semantic segmentation (assigning a class label to each pixel in an image). The results showed that fine-tuning the model on synthetically fog-distorted images improved the model's performance on authentic fog-distorted images, further indicating that synthetic data can be helpful in increasing robustness in adverse weather conditions.

2.3 Image quality assessment

Image quality assessment (IQA) is divided into three fields: full-reference (FR-IQA), reduced-reference (RR-IQA), and no-reference (NR-IQA) [21, 22]. FR-IQA refers to methods where an image's quality is determined by comparing it to a reference image with assumed perfect quality. RR-IQA works by comparing the image to a set of extracted features from the perfect reference image. Finally, NR-IQA refers to methods where the evaluated

images are assessed independently of any reference information. Since images captured by cameras on AVs lack corresponding reference data, this thesis will only consider NR-IQA methods.

Methods of NR-IQA can be classified into distortion-specific and general-purpose methods [23]. Distortion-specific methods are designed to evaluate the quality of images affected by known distortions. Examples of such distortions include JPEG and JPEG2000 compression, for which several distortion-specific NR-IQA methods have been developed [24]. In contrast, general-purpose methods are designed to evaluate the quality of images regardless of the type of distortion. As weather-related noise comes in multiple forms (rain and fog), only general-purpose methods will be considered in this thesis.

For the purpose of comparison, this thesis organizes NR-IQA methods into five overarching paradigms: NSS, CNN, attention-based, CLIP-based, and LMM. These categories are not standardized in the literature nor exhaustive, as some methods span multiple categories. Nevertheless, these paradigms do reflect methodological distinctions found across NR-IQA research. The following sections introduce each paradigm in turn, along with the specific NR-IQA and NR-PCQA methods used in the evaluation.

2.3.1 Natural scene statistics

A prominent research area within general-purpose NR-IQA is the field of natural scene statistics (NSS). NR-IQA based on NSS relies on the insight that high-quality natural images exhibit consistent statistical properties [25] and distorted images typically possess measurable deviations from these properties. Using these insights, Mittal et al. [26] developed BRISQUE, an NSS-based NR-IQA method in the spatial domain. It computes mean-subtracted contrast-normalized (MSCN) coefficients and fits them to a generalized Gaussian distribution model. Additionally, BRISQUE computes pairwise products of neighboring MSCN coefficients using an asymmetric generalized Gaussian distribution model. The resulting features are then fed into a support vector machine, which predicts the scores.

While BRISQUE outperforms earlier NR-IQA methods, one limitation is that it is an opinion-aware model, meaning it is trained on a dataset of images with corresponding subjective scores, which are scores provided by humans [27]. A disadvantage of this approach is that it is costly and time-consuming to collect images with subjective scores. For instance, Ribeiro et al. [28] found that accurate subjective scores of an image cost approximately

0.60 $. Although this may appear cheap, the total cost becomes significant when considering that NR-IQA datasets often comprise tens of thousands of images. Furthermore, opinion-aware models may struggle to detect distortions in images that do not appear in their training data. To address these issues, Mittal et al. [27] introduced NIQE, an NSS-based NR-IQA method similar to BRISQUE that does not rely on human-labeled training data. This makes NIQE opinion-unaware, meaning that it does not rely on subjective scores of images. Whereas BRISQUE is trained on features extracted from both natural and distorted images along with human opinion scores, NIQE is exclusively trained on the NSS features of natural high-quality images. These features are fitted to a multivariate Gaussian (MVG) model. Evaluation is performed by calculating the distance between the MVG fitted to the NSS features of the test image and the MVG of the features of the natural high-quality images.

Extending this approach, Zhang et al. [29] introduced IL-NIQE, which enhances NIQE by introducing three additional NSS features: quality-aware gradient features, color distortions, and log-Gabor filters, which detect patterns at different scales and orientations. IL-NIQE also computes local quality scores for each image patch and then averages them across all patches.

2.3.2 Deep learning approaches

One consistent limitation of NSS-based approaches is the difficulty of finding representations that accurately model complex images and distortion types [30]. The advantage of a deep learning approach is that the network automatically learns optimized feature representations without human intervention. This makes deep learning approaches more robust, and as a result, they typically outperform handcrafted approaches, such as NSS [31].

Zhang et al. [32] proposed DBCNN, a deep learning-based NR-IQA model comprising two convolutional neural networks (CNNs): one for synthetic distortions and the other for authentic distortions. The synthetic-distortion CNN is a modified version of VGG-16, a widely used deep CNN architecture, trained on a large-scale dataset of synthetically distorted images from the Waterloo Exploration and PASCAL VOC datasets. The CNN aimed at authentic distortions is also based on VGG-16, but instead leverages features learned from the ImageNet database. The output of both networks is combined via bilinear pooling to form one representation, which is used for quality prediction. The proposed approach achieves competitive performance, outperforming several NSS-based and deep learning-based NR-IQA methods.

With the rise of transformers and their defining feature, self-attention,

several NR-IQA models have begun to incorporate attention mechanisms to improve performance. Chen et al. [33] proposed TOPIQ, a top-down network that emulates the human visual system by using self-attention to steer the network towards relevant parts of an image. Whereas traditional bottom-up approaches start with low-level features and build up to high-level ones, TOPIQ begins with high-level semantic features and progressively guides lower-level features through downward propagation. Additionally, TOPIQ employs cross-scale attention, enabling layers responsible for fine-grained detail to incorporate information from layers that capture higher-level features. The authors write that this unique architecture achieves competitive performance, outperforming all CNN-based methods included in the comparison, as well as MUSIQ and TReS, two prominent transformer-based NR-IQA models.

Other deep learning models utilize CLIP, a neural network trained to embed images and text in the same embedding space [34]. It was trained using 400 million image-text pairs collected from the internet and consists of two encoders: one for images and one for text. The training objective is to maximize the cosine similarity between true image-text pairs and minimize the cosine similarity for mismatched pairs. Using CLIP, Agnolucci et al. [35] proposed QualiCLIP, a self-supervised opinion-unaware NR-IQA approach. This offers two key advantages: it removes the need for costly human annotations and generally improves generalization across datasets. Instead of relying on subjective scores, QualiCLIP is trained to rank synthetically degraded images according to their similarity in CLIP's feature space to antonym prompts such as "Good photo" and "Bad photo". Furthermore, it is trained such that images with similar content consistently obtain similar feature representations. The authors write that this approach yields an NR-IQA method that achieves state-of-the-art performance among opinion-unaware approaches and even surpasses many opinion-aware methods.

A recent research trend involves applying large multimodal models (LMMs) to the problem of NR-IQA, using LMMs' ability to capture semantic relationships between images and text. Wu et al. [36] proposed a methodology for teaching LMMs to assess image quality by aligning image representations with text embeddings of discrete quality levels. Their approach is motivated by the observation that LMMs, like humans, typically assess images with qualitative descriptions rather than numerical scores. Thus, the authors converted the mean opinion scores into five quality labels: "bad," "poor," "fair," "good," and "excellent." Using this approach, the authors extended mPLUG-Owl-2, a multi-modal model which maps vision and text through

modality-aware learning, and developed Q-Align, an LMM with state-of-the-art performance on NR-IQA benchmarks [36].

2.4 No-reference point cloud quality assessment

Similar to image quality assessment, point cloud quality assessment is divided into three fields based on the availability of information during the evaluation. These are full-reference (FR-PCQA), reduced-reference (RR-PCQA), and no-reference (NR-PCQA) methods. Since LiDAR point clouds collected by AVs typically lack reference data, only NR-PCQA models will be considered in this thesis. These methods are typically classified into three types: model-based, projection-based, and hybrid approaches combining both. The following sections introduce each type in turn.

2.4.1 Model-based NR-PCQA

Model-based NR-PCQA techniques extract color and geometry information directly from 3D point clouds without projecting them onto a 2D plane [37]. Liu et al [38] proposed ResSCNN, a model-based NR-PCQA method using sparse convolutional layers. ResSCNN consists of three components: hierarchical feature extraction, pooling and concatenation, and score prediction. The hierarchical feature extraction uses sparse convolutional layers, which are similar to standard convolutional layers but only operate on the non-zero and non-missing elements of the input data [39]. After the feature extraction, the network pools and concatenates the feature vectors, and the score is predicted using a module consisting of two fully connected layers. ResSCNN achieved state-of-the-art performance on NR-PCQA benchmarks, outperforming both FR-PCQA and other NR-PCQA methods [38].

In addition to using deep learning, some NR-PCQA techniques draw inspiration from image quality assessment and utilize NSS to process point clouds. Zhang et al. [40] introduced 3D-NSS, which uses NSS computed from geometric and color properties of a point cloud. Specifically, 3D-NSS estimates curvature, anisotropy (a measure of variation of geometrical properties depending on direction), linearity, planarity, and sphericity of each point. These features are then aggregated, and the resulting vector is fed to a support vector regressor that predicts the quality of the point cloud.

2.4.2 Projection-based NR-PCQA

Projection-based methods for NR-PCQA differ from model-based ones in that they begin by projecting the points in the point cloud onto one or more 2D planes. The quality evaluation is then carried out on these projections, rather than the raw point cloud data. The advantage of this approach is that it is more computationally efficient and leverages existing techniques for assessing image quality.

Liu et al. [41] introduced PQA-Net, one of the first no-reference methods for point cloud quality assessment. The model projects the point cloud onto six 2D image planes and then extracts features using a CNN. The resulting feature vector is then fed into a distortion classifier and a quality predictor, whose outputs are combined via inner product to produce the final score. Chai and Shao [42] introduced MS-PCQE, a projection-based NR-PCQA technique that uses projections of two different focal lengths. For each focal length, the model projects the point cloud onto six planes (front, back, left, right, up, and down). These projected images are then processed by a residual neural network to generate feature maps, which are subsequently fed into a ConvGRU module. The output is then passed through a mask-aware transformer block with multi-head attention. This process is repeated four times before the final output is processed by a multi-layer perceptron to predict the score of the point cloud.

2.4.3 Hybrid methods

Hybrid NR-PCQA methods combine both projection-based and model-based techniques. The purpose of this multi-modal approach is to capture distortions that a single modality cannot. For instance, Zhang et al. [43] note that structural degradation and geometry downsampling are more easily detected in the point cloud modality, while color quantization and noise are better detected in the image modality.

Zhang et al. [43] introduced MM-PCQA, the first multi-modal approach to NR-PCQA. MM-PCQA works by splitting the point cloud into sub-models and then using a point cloud encoder to create quality-aware embeddings. The projected images are rendered directly and encoded using an image encoder. The point cloud and image embeddings are then fused using a symmetric cross-modality attention block. The results of this block are then fed to a quality regression module. The authors emphasize that this multi-modal technique outperformed all state-of-the-art NR-PCQA techniques it was tested against.

2.5 Evaluation methods

NR-QA methods, used here as a term to encompass both NR-IQA and NR-PCQA methods, are typically assessed by comparing their predicted scores with human evaluations of the same images or point clouds. Human evaluations, also known as subjective scores, are gathered by asking humans to score images or point clouds, typically on a five-point scale [44, 37].

Using the subjective scores, NR-QA methods are typically assessed using the following four metrics: Pearson linear correlation coefficient (PLCC), Spearman's rank correlation coefficient (SRCC), Kendall rank correlation coefficient (KRCC), and Root mean square error (RMSE) [25, 37, 23]. PLCC and RMSE assess the accuracy of predicted scores by measuring their deviation from corresponding subjective scores. In contrast, SRCC and KRCC evaluate how well the predicted scores preserve the rank order of the subjective evaluations.

SRCC measures the prediction monotonicity of the NR-QA by comparing the predicted scores with the ground-truth quality ranks. The formula is expressed as:

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{2.1}$$

where d_i is the difference between the objective and ground-truth quality rankings for image or point cloud i.

To illustrate the use of SRCC, consider the following example. Suppose a point cloud has been distorted four times, with each version having an increasing level of noise. These versions can be ranked in order of increasing distortion as $O = [1, 2, 3, 4]$. An NR-PCQA method then scores each of the point cloud versions, and the scores are ranked from best to worst. This results in the ranking $P = [1, 3, 4, 2]$. Calculating the SRCC is then done using the following calculation:

$$\text{SRCC} = 1 - \frac{6((O_1 - P_1)^2 + (O_2 - P_2)^2 + ... + (O_n - P_n)^2)}{N(N^2-1)} \tag{2.2}$$

$$= 1 - \frac{6((1-1)^2 + (2-3)^2 + (3-4)^2 + (4-2)^2)}{4(4^2-1)} = 0.4 \tag{2.3}$$

KRCC, like SRCC, uses ranks as opposed to absolute scores in its evaluation [45]. 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 $Y = \{Y_1, ..., Y_n\}$, we consider all pairs $\{(i, j) : i <$

$j \leq n\}$. A pair is concordant, discordant, or tied depending on the following conditions:

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

With these definitions, the formula of KRCC (tau-b) is as follows:

$$\tau = \frac{C - D}{\sqrt{(C + D + T_X) * (C + D + T_Y)}} \tag{2.4}$$

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.

Using the same example as before, with ground-truth ranking of $[1, 2, 3, 4]$ and predicted rankings of $[1, 3, 4, 2]$. The concordant pairs (by index) are $[(0, 1), (0, 2), (0, 3), (1, 2)]$, the discordant ones are $[(1, 3), (2, 3)]$, and there are no ties. This means that the KRCC is calculated as:

$$\text{KRCC} = \frac{4 - 2}{\sqrt{(4 + 2 + 0) * (4 + 2 + 0)}} = \frac{1}{3}$$

2.6 Related work

Previous work on evaluating NR-IQA models has typically involved benchmark image datasets with subjective scores. The image benchmark datasets are typically divided into two categories: synthetic and authentic datasets. Synthetic datasets consist of image datasets where high-quality images have been distorted by applying artificial degradations, such as blur, compression, or other forms of noise. Authentic image datasets consist of images captured in real-world conditions where images are distorted due to factors such as poor lighting, motion, or lens blockages.

Sheikh et al. [46] developed the LIVE dataset, one of the first and most widely used synthetic benchmarks for NR-IQA. It contains 779 distorted images that were generated by distorting 29 pristine reference images. Five distortion types were used in the process: JPEG compression, JPEG2000 compression, white noise, Gaussian blur, and fast fading. Despite having a limited range of distortion types, LIVE remains a standard benchmark in the field of NR-IQA. Another prominent synthetic benchmark developed by Ponomarenko et al. [47] is TID2013, which comprises 3000 distorted images generated from 25 reference images using 25 different distortion

types. In addition to standard degradations such as Gaussian noise and JPEG compression, TID2013 includes several more specialized distortion types, such as chromatic aberrations, mean shift, and non-eccentricity pattern noise.

Ghadiyaram and Bovik [48] introduced CLIVE, a widely used NR-IQA benchmark dataset containing authentic distortions of images. It consists of 1162 images captured under natural conditions using a variety of mobile phones. As a result, the images in CLIVE have complex and varied real-world distortions stemming from factors such as low-light conditions, defocus, low resolution, or motion blur. The authors obtained subjective quality scores through a large-scale crowdsourced study involving over 8000 participants, which provided robust ground-truth values.

Using these established benchmark image datasets, several studies have evaluated the performance of NR-IQA models. Xu et al. [45] evaluated several NSS-based NR-IQA methods on three benchmark datasets: LIVE, TID2013, and CSIQ. Among the evaluated methods, IL-NIQE and GMLOGQA, an NSS method based on gradient magnitude and Laplacian features, demonstrated the best overall performance across the three datasets. However, the authors also observed that despite their strong average performance, these methods struggled to accurately assess images with noise distortions that were not present in their training data.

Yang et al. [30] advanced this research by evaluating the performance of deep neural network-based methods for NR-IQA. They benchmarked these models on four synthetic datasets: LIVE, TID2013, CSIQ, LIVE MD, as well as on the authentic CLIVE database. The authors found that deep learning–based approaches generally outperformed NSS-based methods, attributing this to the ability of deep models to automatically learn optimized feature representations. Among the compared methods, the authors found that DBCNN achieved excellent performance, outperforming NSS-based methods such as IL-NIQE and BRISQUE.

Many papers proposing NR-IQA methods also include comparisons against existing approaches using benchmark image datasets. In their paper introducing Q-Align, Wu et al. [36] compare it against several state-of-the-art NR-IQA methods. When trained on the authentic SPAQ dataset, Q-Align outperformed all six other NR-IQA methods across five benchmark datasets. These included the CNN-based methods DBCNN, NIMA, and HyperIQA, along with the CLIP-based approaches CLIP-IQA+ and LIQE, as well as the transformer-based method MUSIQ.

Similar to NR-IQA, previous work in NR-PCQA consists of benchmark point cloud datasets with subjective scores. Yang et al. [49] introduced SJTU-

PCQA, an NR-PCQA benchmark comprising 378 synthetically distorted point clouds, generated from 10 reference point clouds using 7 distortion types. The distortion types include color and geometry noise, as well as noise due to scaling and compression. Another widely used NR-PCQA benchmark is the Waterloo Point Cloud (WPC) dataset, developed by Liu et al. [50, 51], which consists of 20 high-quality point clouds and 740 distorted ones. Five distortion types were used, including downsampling, Gaussian noise, and compression artifacts resulting from three algorithms for point cloud compression.

Several evaluations of NR-PCQA methods have been performed on these NR-PCQA benchmarks. Zhou et al. [22] surveyed the area of PCQA techniques, looking at model-based and projection-based methods and their performance on four datasets. While the review is not specific to NR-PCQA, it benchmarked several NR-PCQA techniques on the STJU-PCQA and WPC datasets. The results show that MM-PCQA outperforms all other NR-PCQA methods in the comparison across both datasets. Porcu et al. [37] surveyed the field of NR-PCQA directly. The authors review the state-of-the-art in NR-PCQA concerning both techniques and datasets. The authors do not directly compare the performance of NR-PCQA methods but instead use the results of Chai and Shao [42], who introduced MS-PCQE and evaluated it against five other NR-PCQA models. The results of the comparison show that MM-PCQA and MS-PCQE achieved the highest performance across the five point cloud datasets used in the study.

This thesis distinguishes itself from prior work by focusing on the evaluation of NR-IQA and NR-PCQA methods in the realm of autonomous driving. While previous studies assess models on general-purpose image and point cloud benchmarks, this work is limited to images and point clouds that are captured in an autonomous driving context.

Chapter 3

Method

3.1 Literature search

A comprehensive literature review is carried out to gain a clear understanding of the thesis's core topics. The key search words were NR-IQA, NR-PCQA, image distortion, and point cloud distortion. The primary tool used for finding literature is Google Scholar due to its broad coverage of scientific papers and journals. In the literature search, both survey articles and original research papers are considered. Survey articles are primarily used to gain a broad understanding of the fields of NR-IQA and NR-PCQA, while original research papers provide detailed information on specific methods. Both types of articles are also used to compare the performance of NR-IQA and NR-PCQA techniques. During the literature search, factors such as citation count, publication venue, year, and language were considered to ensure the quality of the chosen articles.

3.2 Evaluation approach

The first considered evaluation approach for the NR-IQA and NR-PCQA methods was to evaluate them on authentic weather-distorted images and point clouds. However, the problem with this approach is that no dataset containing either weather-distorted images or point clouds with corresponding subjective scores could be found. Notably, no major survey on NR-IQA or NR-PCQA mentions the existence of a dataset tailored explicitly to the AV domain [25, 30, 37, 22]. Pursuing this option would therefore require conducting a large-scale survey with human participants to obtain subjective scores. Given the substantial effort required, this method was ultimately rejected.

(a) OD confidence on the two leftmost bounding boxes is 0.78 and 0.77

(b) OD confidence on the two leftmost bounding boxes is 0.85 and 0.87

Figure 3.1: Example where object detection (OD) confidence does not correlate with differences in image quality. Images are taken from the AVL dataset [7]

.

A second approach, considered only for NR-IQA evaluation, involved using object detection performance as a proxy for ground-truth image quality. This method relies on the intuition that since images in AVs are ultimately used in object detection algorithms, it is logical to assess image quality based on the confidence of the object detection algorithm.

The problem with this approach, however, is that object detection performance is not always correlated with image quality. For instance, in Figure 3.1, the two images are captured within a very short interval and appear to have precisely the same quality. However, because the vehicles on the right side of the scene are positioned closer to the camera in the second image, the object detection algorithm assigns higher confidence scores compared to the first image. Specifically, it assigns 0.85 and 0.87 in the second image and 0.78 and 0.77 in the first. These confidence scores represent the model's estimated probabilities that the detected objects are cars. If the ground truth was set based on these scores, the second image would be assigned a higher quality score than the first, despite the two images exhibiting identical quality. Thus, object detection confidence was deemed unreliable in setting ground truth scores, and this approach was rejected. Chapter 6.1 considers alternative methods for approaching this problem.

The third approach relies on the insight that synthetic noise distortions at various levels can generate datasets with known quality rankings. To accomplish this, a reference image or point cloud is used and then distorted several times with increasing levels of noise. This produces a set of images or point clouds where the absolute quality of the samples is unknown, but the relative quality rankings are known. This approach was ultimately the one adopted in the thesis.

Building on this approach, the evaluation of an NR-QA method is as follows. An image or point cloud set with known quality rankings is generated using the method described above. The NR-QA method then scores all samples in the set, generating a list of scores. The performance of the NR-QA is then evaluated by comparing its rankings with the ground-truth rankings using Spearman rank correlation coefficient (SRCC) and Kendall rank correlation coefficient (KRCC). Since there are no subjective scores for the images or point clouds, PLCC and RMSE cannot be used as evaluation metrics.

The approach of applying synthetic distortions to generate image sets with known quality rankings has previously been used in the field of NR-IQA. Liu et al. [52] utilized it to generate RankIQA, an NR-IQA method trained to rank image pairs degraded by four distortion types. Additionally, Agnolucci

et al. [35] synthetically distorted images with increasing amounts of noise and trained CLIP to rank the images according to distortion severity.

The approach in this thesis differs from that of Liu et al. and Agnolucci et al. in a few distinct ways. First, it utilizes larger image sets with known relative quality rankings, rather than the small sets of synthetically degraded images employed by both Liu et al. and Agnolucci et al. Secondly, whereas Liu et al. and Agnnolucci et al. use conventional image distortions such as GB, GN, JPEG, and JP2K, the synthetic distortions in this work will include weather-related distortions. Lastly, while previous work has used the approach to train an NR-IQA method, this thesis employs it to evaluate NR-IQA and NR-PCQA methods.

3.3 Data collection

The thesis utilizes two datasets for the evaluation: the Finnish Geospatial Institute (FGI) dataset [53] and the REHEARSE dataset [54]. Each dataset was chosen for different parts of the evaluation.

The FGI dataset [53] includes data from two drives: one urban drive around Otaniemi, Espoo, and another covering both urban and rural scenery, extending from Espoo to Munkkivuori in Helsinki. These two drives are referred to as Otaniemi and Munkkivuori, respectively. The Otaniemi drive was captured during the day, resulting in brighter images, whereas the Munkkivuori drive was captured in the evening, resulting in darker images. Both drives were recorded on December 3, 2023, during winter conditions and include data from a LiDAR, an RGB camera, and a thermal camera. The FGI dataset is stored in ICE S3 storage, a datacenter operated by RISE, and retrieved using the `boto3` API. The Munkkivuori drive consists of 6915 PNG images, while the Otaniemi drive consists of 7851 PNG images.

The REHEARSE dataset [54] comprises LiDAR, image, and RADAR data collected from an outdoor test track and an indoor tunnel using static sensors. The data is collected under clear, rainy, and foggy conditions, with the rain and fog artificially generated using sprinklers. Like the FGI dataset, the data in the REHEARSE dataset is stored in ICE S3 Storage and was retrieved using the boto3 API. The point clouds are stored as binary files, each containing a sequence of points of type (x, y, z, i) where (x, y, z) are the Cartesian coordinates and i is the intensity value.

The FGI dataset was selected for the NR-IQA evaluation due to its authentic conditions, providing realistic winter road scenes encountered during real-world driving. Although the REHEARSE dataset contains image

data, it is only images captured on a test track or in a tunnel, not in authentic driving conditions. From each drive of the FGI dataset, 10 images were selected as reference images for the synthetic distortions using Python's standard random module with a fixed seed to ensure reproducibility. This results in a total of 20 reference images. The number of images was chosen to balance scene diversity with computational efficiency. The selected reference images can be seen in Figure 3.2.

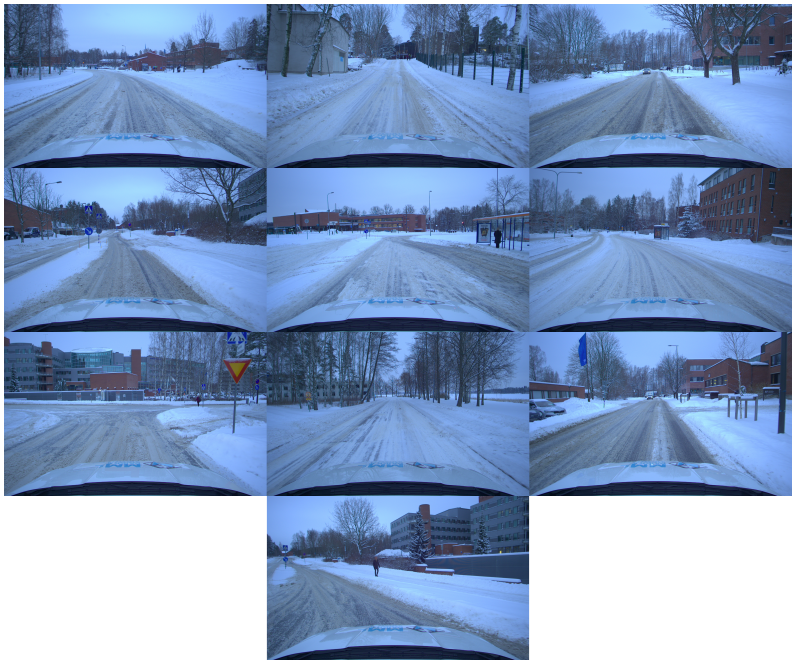
For the point cloud data, the REHEARSE dataset was selected over the FGI dataset. The reason is that initial experiments in the project applying NR-PCQA methods to the FGI point clouds yielded poor results. The REHEARSE dataset [54], recorded under more controlled conditions, was therefore used instead. Only point clouds from the outdoor test track were used. The indoor tunnel data were excluded because extensive preprocessing was required to make the data compatible with the NR-PCQA evaluation. A total of 40 point clouds were randomly selected from the outdoor test track, comprising 20 captured in clear weather and 20 in rainy conditions. Of these, 27 were recorded during the day, and 13 were recorded during the night. The number of reference point clouds was doubled compared to the images, as only one distortion type was used in the NR-PCQA evaluation (see Section 3.4)

3.4 Data preprocessing

3.4.1 Image preprocessing

The `Albumentations` Python library [16] is used to synthetically apply weather distortions to images. `Albumentations` is an image augmentation library featuring over 100 transformations for augmenting images. The selection of `Albumentations` is motivated by its support for weather-related distortions, particularly artificial rain and fog, as well as its ability to generate image sets with monotonically increasing noise severities. Furthermore, `Albumentations` does not require depth information, which is not available for the images and is often a prerequisite for many synthetic distortion techniques. `Albumentations` has also been successfully used as a tool for applying synthetic weather distortions in prior studies [55, 56, 57]. The thesis uses two types of synthetic weather distortions for images: rain and fog. Noise from falling snow was excluded because Albumentations does not implement it, and no suitable library capable of generating image sets with monotonically increasing snow noise was found.

Within `Albumentations`, the functional transform `add_rain` is used



(a) The ten reference images used from the Otaniemi drive.

(b) The ten reference images used from the Munkkivuori drive.

Figure 3.2: The twenty reference images used from the Otaniemi and Munkkivuori drives.

to create synthetic images with rain effects. The `add_rain` transform adds semi-transparent streaks to an image, simulating the appearance of raindrops. This is achieved by first defining a set of coordinates as the starting points for the raindrops, along with their slant angle, width, length, and color. Using this information, `add_rain` utilizes `OpenCV` to draw a line for each specified raindrop. `add_rain` also blurs the image using a box blur with a kernel size determined by the `blur_value` parameter. Finally, it changes the brightness of the image by converting it to HSV format and multiplying the value component by the `brightness_coefficient` parameter.

For each image selected in Chapter 3.3, 100 versions with synthetic rain are generated using the parameters outlined in Table 3.1. This builds upon earlier work of synthetic image distortion, notably that of Ahar et al. [58], which utilized 100 distortion levels to capture subtle differences in image quality.

Parameters listed as constants remain the same for all versions of the image. For parameters listed as ranges $[a, b]$, the values are linearly spaced across the 100 versions. Thus, the first version has a value of a, the last version has a value of b, and the step size is given by $(|b - a|/(n - 1)$. To simulate a gradual and uniform increase in rain-induced distortion, `droplet_share` is progressively increased to simulate denser rainfall, while `brightness_coefficient` is decreased to reflect reduced visibility due to the intensified rain. `slant`, `drop_length`, `drop_width`, and `blur_value` are kept constant to maintain consistent raindrop geometry and avoid abrupt visual changes. An example of three versions of the same image with varying amounts of rain distortion is shown in Figure 3.3.

Fog effects are synthesized using the `add_fog` transform from `Albumentations`. `add_fog` simulates fog by taking an input image, a fog intensity coefficient (`fog_intensity`), a transparency value (`alpha_coef`) for the fog particles, a list of coordinates specifying their positions, and a list of their respective radii. For each specified fog particle, `add_fog` uses `OpenCV` to draw a circle on a copy of the image. The image with the circle is then blended with the original image using the following formula:

$$S(i, j, c) = \alpha F(i, j, c) + (1 - \alpha)S(i, j, c)$$

where $S(i, j, c)$ refers to the pixel value at location (i, j) and color channel c in the original image, and $F(i, j, c)$ refers to the corresponding pixel in the fog circle. The blending factor α is given by the formula $\alpha =$ alpha_coef $\times$ fog_intensity.

Table 3.2 shows the parameters used to add synthetic fog noise to the

Parameter name	Explanation	Value
`droplet_share`	Proportion of pixels where a droplet is placed. Since droplets are larger than a single pixel, the total share of affected pixels is greater.	$[0.0, 0.01]$
`rain_drops`	Array of pixel coordinates indicating where droplets are placed. The droplet placement is randomized (seeded for reproducibility), and the number of raindrops is controlled by `droplet_share`.	
`slant`	The angle in degrees at which the synthetic raindrops are applied.	15
`drop_length`	Length of each raindrop in pixels.	20
`drop_width`	Width of each raindrop in pixels.	1
`blur_value`	The value N used in an NxN box filter applied to the image. This is done since rainy images are typically blurrier [16].	7
`brightness_coeff`	Coefficient applied to each pixel value to reduce brightness, as rainy images are typically darker.	$[1.0, 0.7]$

Table 3.1: Parameters for the synthetic rain distortion

images. For each reference image across both locations, 100 versions with synthetic fog are synthesized using the parameters in Table 3.2. The progressive increase is linearly spaced, following the same approach as for the rain-distorted images described earlier. The choice of 100 distortion levels (as with the rain-distorted images) follows the reasoning of Ahar et al. [58] to capture subtle image quality differences. An example of an image distorted three times with increasing fog severity is shown in Figure 3.4.

(a) droplet_share = 0.003, brightness_coefficient = 0.924

(b) droplet_share = 0.005, brightness_coefficient = 0.848

(c) droplet_share = 0.007, brightness_coefficient = 0.788

Figure 3.3: The same image from Otaniemi with different amounts of rain-related noise.

(a) $\alpha = 0.176$

(b) $\alpha = 0.354$

(c) $\alpha = 0.531$

Figure 3.4: The same image from Munkkivuori with different increasing amounts of foggy noise.

3.4.2 Point cloud preprocessing

The work of Hahner et al. [19], referred to as `LiDAR_fog_sim`, is a tool for synthetically distorting point clouds. In this study, `LiDAR_fog_sim` was selected to synthetically distort the point clouds. The choice to use `LiDAR_fog_sim` for point cloud distortion is motivated by two factors. Firstly, it is a state-of-the-art tool that simulates fog in a physically accurate manner [19]. Secondly, it has been used in prior research to generate synthetic fog distortions in point clouds [59, 60].

`Lidar_fog_sim` simulates fog by extending an optical LiDAR model proposed by Rasshofer et al. [61]. The algorithm Hahner et al. use to distort fog particles is outlined in Figure 3.5. For each point p, the algorithm inputs the point's intensity i, differential reflectivity β_0, and the half-power pulse width τ_H. The differential reflectivity models how much of the LiDAR pulse is reflected back by the target object. The half-power pulse width is a parameter that defines the duration over which a LiDAR pulse maintains significant power around its peak.

The algorithm starts by calculating the distance R_0 to the point. Lines 3-5 compute the response intensity from the hard target i_{hard}, factoring in the attenuation coefficient α. The hard target refers to the object or surface that reflects the LiDAR pulse to the sensor. Lines 6-8 numerically compute integrals used to calculate the intensity from the soft target, which refers to the fog particles. This is done as the expression does not possess a closed form and is done for all distances $R \in \{0.1, ..., R_0\}$.

Next, lines 9-11 compute the intensity from the soft target by taking the largest value from the precomputed integrals i_{tmp} and multiplying that with C_A, P_0, and β. C_A is a system constant, P_0 the pulse's peak power, and β is the backscattering coefficient. On line 12, the algorithm checks if the intensity of the soft target exceeds the intensity of the hard target. If this is the case, then the point is shifted by a scaling factor $s = R_{\text{tmp}}/R_0$ and a random noise factor, and its intensity is set to that of the soft target. Otherwise, the point's coordinates remain intact while the point's intensity value is changed to i_{hard}.

Algorithm 1 LiDAR fog simulation

procedure FOGGIFY($\mathbf{p}$, i, α, β, β_0, τ_H)
  $R_0 \leftarrow \|\mathbf{p}\|$
  $x, y, z \leftarrow \mathbf{p}$ ▷ $i = P_{R,\text{clear}}$
  $C_A P_0 \leftarrow i \frac{{R_0}^2}{\beta_0}$ ▷ follows from Eq. (12)
  $i_{\text{hard}} \leftarrow i \times \exp(-2\alpha R_0)$ ▷ see Eq. (17)
  for R in $(0, 0.1, ..., R_0)$ **do** ▷ 10cm accuracy
    $I_R \leftarrow \text{SIMPSON}(I(R, R_0, \alpha, \tau_H))$ ▷ see Eq. (18)
  end for
  $i_{\text{tmp}} \leftarrow \max(I_R)$
  $R_{\text{tmp}} \leftarrow \arg\max(I_R)$
  $i_{\text{soft}} \leftarrow C_A P_0 \, \beta \times i_{\text{tmp}}$ ▷ see again Eq. (18)
  if $i_{\text{soft}} > i_{\text{hard}}$ **then**
    $s \leftarrow \frac{R_{\text{tmp}}}{R_0}$ ▷ scaling factor s
    $p \leftarrow \text{RANDOM\_UNIFORM\_FLOAT}(-1, 1)$
    $n \leftarrow 2^p$ ▷ noise factor $n \in (\frac{1}{2}, 2)$
    $x \leftarrow s \times n \times x$
    $y \leftarrow s \times n \times y$
    $z \leftarrow s \times n \times z$
    $i \leftarrow i_{\text{soft}}$
  else ▷ keep original location
    $i \leftarrow i_{\text{hard}}$ ▷ only modify intensity
  end if
  return x, y, z, i
end procedure

Figure 3.5: The algorithm used to apply fog distortion to a particle p with intensity i [19]

Parameter	**Explanation**	**Value(s)**
`fog_particle_share`	Proportion of pixels where a fog particle is placed. As particles are larger than a single pixel, the total proportion of pixels is greater.	0.5
α	Factor determining the visual strength of the fog applied to the image, between 0 and 1.	$[0.0, 0.7]$
`fog_particle_size`	Size of each fog particle in pixels.	25
`fog_particle_positions`	Coordinates where fog particles are placed. Randomized and seeded for reproducibility.	5
`fog_particle_radiuses`	Radius of fog particles in pixels.	25

Table 3.2: Parameter values used for synthetic fog distortion.

Adding synthetic fog to point clouds first involves precomputing the integrals described on line 7 in Figure 3.5. The integrals are computed using Simpson's $1/3$ rule, a numerical integration method using quadratic interpolation. The step size is 0.1 m, $\tau_h = 20$ ns, and the maximum distance $R_0 = 200$ m. Each of the precomputed integral values is the same as in the original publication describing the method [19]. The precomputed integrals are subsequently used to add synthetic fog distortions to the point clouds.

Each of the selected reference point clouds is distorted at ten distortion levels with the parameters described in Table 3.3. The choice of ten distortion levels was motivated by two factors. First, it became increasingly challenging to distinguish perceptual quality differences in point clouds with more than ten levels of distortion. Additionally, ten distortion levels have been used in prior work on point cloud quality degradation [62].

For each point in each point cloud, the algorithm shown in Figure 3.5 is applied. All parameters, except α, are kept consistent with the original paper to ensure alignment with the proposed fog simulation model. α is progressively increased to simulate denser fog in the same linear manner used in the synthetic image distortions. A visualization of a point cloud distorted using three different α values can be seen in Figure 3.6.

3.4.2.1 Color inference of point clouds

The NR-PCQA techniques used in the thesis, which are described in Chapter 3.6, expect point clouds of the form $P = \{g_m, c_m\}_{m=1}^{N}$ where $g_m \in \mathbb{R}^{1\times3}$ are the geometric coordinates and $c_m \in \mathbb{R}^{1\times3}$ consists of the RGB color information. As the point clouds in the dataset are of the form $P = \{(x_k, y_k, z_k, i_k)|k = 1, ..., N\}$, the color is inferred. This is done by taking the intensity value i, an integer between 0 and 255, and setting $c_m = (i, i, i)$. This method of assigning the intensity value to all RGB channels has previously been used to infer color from non-colored point clouds [63].

3.5 Selection of NR-IQA techniques

As mentioned under Chapter 2.3, one NR-IQA method from each of the following paradigms was included in the NR-IQA evaluation: NSS, CNN, attention-based, CLIP-based, and LMM. IL-NIQE [29] was selected as the NSS-based approach because of its proven effectiveness on NR-IQA benchmarks and because it builds upon and enhances earlier NSS-based NR-IQA techniques. DBCNN [32] was selected as the CNN-based approach

Parameter name	Explanation	Value
α	Attenuation coefficient which reduces the signal power.	$[0.0, 0.25]$
β	Coefficient for backscattering, which quantifies the portion of the LiDAR signal that is reflected directly back toward the sensor by fog particles. The value of β is the same as in the original paper.	$0.046\alpha/\ln(20)$
β_0	Differential reflectivity of the target.	$10^{-6}/\pi$
τ_h	Half-power pulse width	20 ns

Table 3.3: Parameters for the synthetic fog distortion of point clouds.

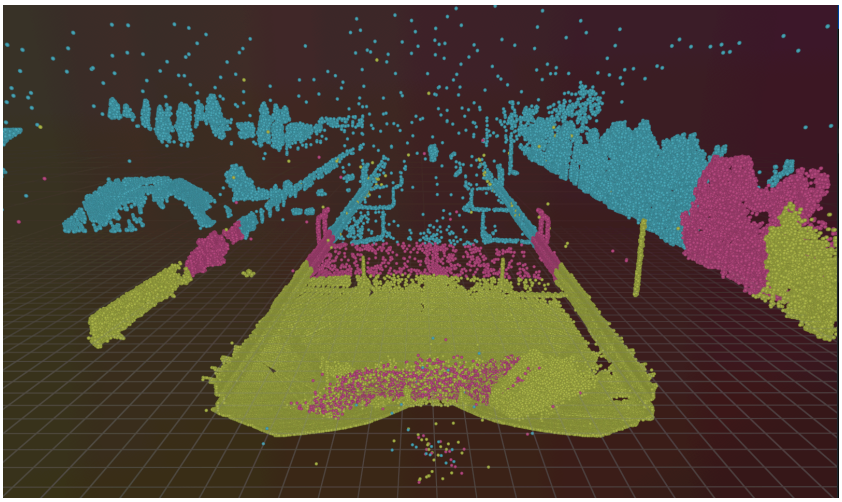


Figure 3.6: Visualization of three point clouds subjected to increasing distortion levels. The blue point cloud corresponds to $\alpha = 0.03$, the pink to $\alpha = 0.06$, and the yellow to $\alpha = 0.09$.

because, as mentioned in Chapter 2.6, it demonstrated the best performance among deep learning methods in a survey of NR-IQA techniques. TOPIQ [33] was selected as the attention-based method. It was chosen because it has demonstrated better performance on NR-IQA benchmarks compared to similar methods, such as TReS [64] and Musiq [65]. QualiCLIP [35] was chosen as the CLIP-based approach as it outperforms other CLIP-based methods such as CLIP-IQA [66] and LIQE [67]. Finally, Q-Align was chosen as the LMM-based model as it has achieved excellent performance on NR-IQA benchmarks [36]. A summary of each of the chosen models can be found in Table 3.4. In this table, the Dataset column refers to the dataset used during training. The dataset training information was ascertained for all methods but IL-NIQE and DBCNN. For all methods except IL-NIQE, a higher score indicates higher image quality.

All of the chosen NR-IQA techniques are sourced from the `pyiqa` library [68], a Python toolbox for image quality assessment that implements both NR and FR-IQA techniques. `pyiqa` is used because it implements all of the selected NR-IQA methods. It also simplifies the evaluation process by eliminating the need to locate and compile individual NR-IQA implementations and has been used in previous NR-IQA research [69, 70].

3.6 Selection of NR-PCQA techniques

This thesis evaluates two NR-PCQA methods, and the reason for limiting the scope to two methods is that each method requires significant configuration and setup. Adding additional methods to the evaluation would be unfeasible given the time constraints.

The selected NR-PCQA methods are MS-PCQE and MM-PCQA. MS-PCQE is selected because it has been shown in previous research to have the best performance on several datasets. Porcu et al. [37] found that MS-PCQE achieved the best performance among all tested NR-PCQA methods on five NR-PCQA datasets. The authors of MS-PCQE developed multiple models, each trained on a different dataset. In this thesis, the model trained on the LS-PCQA dataset [38] is used. LS-PCQA is a large-scale NR-PCQA dataset consisting of 104 reference point clouds, 31 distortion types, and 22568 distorted point clouds. This model is selected because LS-PCQA is by far the largest and most diverse NR-PCQA dataset available [37].

MM-PCQA is chosen since it combines the strengths of projection-based and model-based methods. Additionally, Porcu et al. [37] found that MM-PCQA had similar performance to MS-PCQE and, for some distortion types,

Method	Approach	Dataset	Description
IL-NIQE	NSS	–	NSS approach using a multivariate Gaussian model of image patches, each assessed with a Bhattacharyya-like distance. Final quality score is obtained by average pooling [29].
DBCNN	CNN	–	Deep bilinear model with two CNNs for authentic distortions. Their outputs are pooled for final quality prediction [32].
TOPIQ	Attention-based	KonIQ-10k	Top-down network leveraging high-level semantic features to guide attention to perceptually relevant image regions. Uses coarse-to-fine and cross-scale attention [33].
QualiCLIP	CLIP	KonIQ-10k	Self-supervised method using CLIP to rank synthetically degraded images, removing the need for subjective scores [35].
Q-Align	LMM	KonIQ-10k, SPAQ, KADID-10k	LMM-based model trained with text-defined quality levels rather than numerical scores for better alignment with human ratings [36].

Table 3.4: NR-IQA methods included in the study

even outperformed it. The specific MM-PCQA model used in this thesis is the one trained on the WPC dataset [50, 51]. The model trained on the WPC dataset is selected because it is the only predefined model provided by the authors. While it is possible to train the model on a different dataset, this falls outside the scope of the thesis.

3.7 Analysis

A two-tailed permutation test is conducted for each NR-QA method to evaluate whether its predicted rankings are correlated with the ground-truth rankings. The hypotheses are defined as:

$$H_0 : \text{SRCC}_{\text{mean}} = 0$$
$$H_A : \text{SRCC}_{\text{mean}} \neq 0$$

where $\text{SRCC}_{\text{mean}}$ refers to the mean SRCC value across all image or point cloud sets, depending on whether the method is an NR-IQA or NR-PCQA method. The null hypothesis H_0 states that the mean SRCC is 0, implying that there is no correlation between the NR-QA method's scores and the ground-truth ranking. The alternative hypothesis H_A states that there is a statistically significant correlation.

To estimate the p-value under the assumptions of the null hypothesis, the rankings of the NR-QA method are randomly shuffled within each image or point cloud set. For each permutation, the SRCC is computed using the shuffled rankings and the ground-truth rankings. The resulting SRCCs are then averaged across all sets, resulting in a permuted mean SRCC. This process is repeated 5000 times, and the empirical p-value $\hat{p}$ is computed as:

$$\hat{p} = \frac{1}{5000} \sum_{i=1}^{5000} I(|\text{SRCC}_{\text{mean}}^{(i)}| \geq |\text{SRCC}_{\text{observed}}|)$$

where $\text{SRCC}_{\text{mean}}^{(i)}$ denotes the mean SRCC from the i-th permutation and $\text{SRCC}_{\text{observed}}$ is the mean SRCC obtained from the experimental data. $I()$ is the indicator function that is equal to 1 if $|\text{SRCC}_{\text{mean}}^{(i)}| \geq |\text{SRCC}_{\text{observed}}|$ and 0 otherwise. 5000 is chosen as the number of permutations following the recommendation of Marozzi [71]. As multiple hypotheses are tested, a Holm-Bonferroni correction is performed to minimize the risk of Type 1 errors.

A Friedman test is applied to identify significant differences between two or more of the selected NR-IQA methods. It is a non-parametric statistical test suitable for comparing multiple methods across multiple datasets [72]. In this evaluation, each dataset corresponds to an image set. The null hypothesis states that the methods are equivalent, while the alternative hypothesis states that at least one method differs significantly from at least one other method. The chi-square distribution approximates the Friedman test statistic when the number of methods or image sets is sufficiently large, typically greater than 5 [73]. In the evaluation, five methods and 40 image sets are used, indicating that the chi-square approximation is reasonable.

If the null hypothesis of the Friedman test is rejected, pairwise Wilcoxon signed-rank tests are conducted to identify which NR-IQA methods differ significantly. Additionally, for the NR-PCQA methods, a Wilcoxon signed-rank test is performed to assess whether MM-PCQA and MS-PCQE differ

significantly.

A Wilcoxon signed-rank test including two methods, A and B, is performed by first computing the differences $D = A - B$. The absolute values $|D|$ are then ranked, and zero differences are excluded. Then, the original signs of the differences are reintroduced and summed up into two separate sums. Positive differences are summed up into the sum $\sum R^+$, and negative differences are summed up into the sum $\sum R^-$. A low p-value suggests a statistically significant difference between the methods A and B. If $\sum R^+ > \sum R^-$ and p is sufficiently small, it suggests that A outperforms B [74]. By contrast, if $\sum R^- > \sum R^+$ and p is sufficiently small, it suggests that B outperforms A. Since multiple hypotheses are tested in the pairwise Wilcoxon signed-rank tests for the NR-IQA methods, a Holm-Bonferroni correction is applied to reduce the risk of Type I error.

Chapter 4

Results and analysis

Method	SRCC (SD)	KRCC (SD)
TOPIQ	-0.496** (0.685)	-0.443** (0.622)
DBCNN	0.131** (0.872)	0.160** (0.805)
QualiCLIP	-0.062** (0.756)	-0.070** (0.696)
Q-Align	0.998** (0.002)	0.979** (0.015)
IL-NIQE	0.946** (0.094)	0.868** (0.156)

* $p < 0.05$ ** $p < 0.001$

Table 4.1: Mean SRCC and KRCC values of the NR-IQA techniques across both distortion types and locations, with standard deviations shown in parentheses. The p-values are obtained from permutation tests evaluating whether the mean SRCC and KRCC differ significantly from 0.

4.1 NR-IQA results

With 20 base images selected and two distortion types employed, a total of 40 image sets were generated, resulting in 40 SRCC and KRCC values per NR-IQA technique. Table 4.1 presents the mean SRCC and KRCC values for the NR-IQA techniques under evaluation. Q-Align achieves near-perfect agreement with ground truth rankings, with mean SRCC and KRCC values of 0.998 and 0.970, respectively. In comparison, IL-NIQE also performs well, with corresponding values of 0.946 and 0.868.

In contrast, the performance of TOPIQ, DBCNN, and QualiCLIP is considerably worse. QualiCLIP has a mean SRCC of -0.062 and a standard deviation of 0.756, suggesting that the SRCC values are widely dispersed in

both the positive and negative directions. DBCNN has a higher mean SRCC value of 0.131, but like QualiCLIP, its values vary widely, with a standard deviation of 0.872. TOPIQ has a mean SRCC value of -0.496 and a mean KRCC value of -0.443, indicating a strong negative correlation between its scores and the ground truth scores. In other words, it performs worse than random, consistently assigning higher scores to lower-quality images than to higher-quality ones. TOPIQ also shows high variability, with standard deviations of 0.685 for the mean SRCC and 0.622 for the mean KRCC.

The exact SRCC values of each NR-IQA method on each image set can be seen in Figure 4.1. The points in the figure are grouped based on location (Otaniemi vs Munkkivuori) and distortion (rainy vs foggy). Q-Align has consistently high SRCC scores across the 40 image sets, regardless of distortion type or location. The lowest SRCC observed among these sets is 0.990. IL-NIQE also achieves consistently high SRCC scores, performing slightly better on foggy image sets than rainy ones. On the foggy image sets, IL-NIQE achieves an astonishing mean SRCC of 0.999 with a standard deviation of 0.002. However, on the rainy image sets, the mean SRCC is lower at 0.892 with a standard deviation of 0.110. Thus, IL-NIQE seems to struggle more with rain degradation than with fog degradation. Interestingly, QualiCLIP achieves excellent performance on the fog-distorted images from Otaniemi, achieving a mean SRCC of 0.966. However, on the foggy image sets from Munkkivuori, the opposite is true, as it achieves a mean SRCC of -0.977. Furthermore, the standard deviations of these calculations are pretty minor, at 0.050 and 0.023, indicating that QualiCLIP exhibits a strong consistency in its ranking of foggy images. On the rainy image sets, the results are less consistent, with a mean SRCC value of -0.118 and a relatively high standard deviation of 0.436.

Differences in performance based on location are also evident for TOPIQ, which achieved a mean SRCC of 0.577 on the foggy Otaniemi image sets and -0.928 on the foggy Munkkivuori image sets. The standard deviation is relatively high in both cases. For the Otaniemi image sets, it is 0.494, and for the Munkkivuori drive, it is 0.169. On the rainy image sets, TOPIQ consistently exhibited clear inverse performance, meaning that it rated lower-quality images as having higher quality. It achieved a mean SRCC value of -0.815 and a standard deviation of 0.177.

Inverse performance is also observed for DBCNN on the rainy image sets. It achieved a mean SRCC of -0.601 across all rainy image sets with a high standard deviation of 0.496. However, it achieved much better performance on the foggy image sets across the two locations. On the foggy image sets from

Otaniemi, it achieved a high mean SRCC of 0.998, whereas on the foggy image sets from Munkkivuori, the mean SRCC dropped to 0.730. This indicates that DBCNN, similar to TOPIQ and QualiCLIP, performed worse on images from Munkkivuori than on those from Otaniemi.

Figures 4.2, 4.3, 4.4, 4.5, and 4.6 show boxplots of the scores of the NR-IQA methods. Each Figure is divided into the four combinations of location and weather: Otaniemi-rainy, Otaniemi-foggy, Munkkivuori-rainy, and Munkkivuori-foggy. For each combination, a boxplot is plotted at every fifth image index, using all images with that index in the given combination. The boxplot shows the median as the line inside the box, with the lower and upper edges of the box representing the first and third quartiles, respectively. The whiskers extend to the smallest and largest data points within 1.5 times the interquartile range. The points outside this range are plotted directly.

Figure 4.2 illustrates a general downward trend in DBCNN scores for the foggy image sets from Otaniemi, and to a lesser extent for those from Munkkivuori. For the rainy image sets at both locations, there is a slight upward trend, with SRCC values being negative in 16 out of 20 cases. Furthermore, Figure 4.2 shows a sharp divide between the absolute scores of the foggy and rainy images. All foggy images receive higher scores than all rainy images, showing that DBCNN consistently assesses the foggy images as having better quality.

The scores of TOPIQ are found in Figure 4.3, which displays a slight decreasing trend for the foggy image sets from Otaniemi, with a mean SRCC score of 0.577. In comparison, the performance on foggy image sets from Munkkivuori is significantly worse, with a mean SRCC value of -0.927. This is also evident in Figure 4.3, with the Munkkivuori-foggy plot showing a distinct upward trend. TOPIQ's poor performance on rainy image sets is also evident from the slightly upward trend on the Otaniemi-rainy and Munkkivuori-rainy image sets.

Turning to Figure 4.4, Q-Align demonstrates high consistency in its scoring ability across locations and weather distortions. A uniform downward trend is observed across all conditions, without any outliers. The lowest mean SRCC observed across all subsets occurs for the rainy image sets from Munkkivuori, and even in this instance, Q-Align achieved an astonishing mean SRCC of 0.996. Additionally, the tight spread of the boxplots shows little variability in the scores among images within the same image index.

Figure 4.5 demonstrates that QualiCLIP is inconsistent in its ranking. Figure 4.5 shows a distinct downward trend in scores for Otaneimi-foggy. However, the opposite is true for the Munkkivuori-foggy image sets, where a

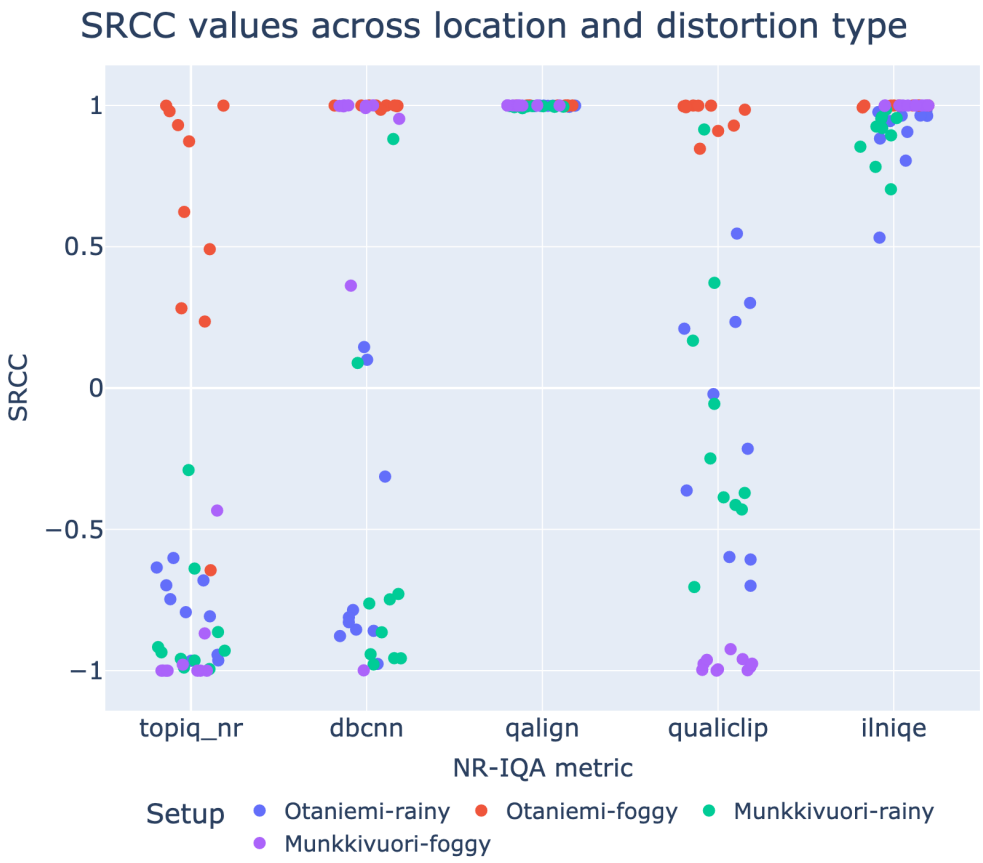


Figure 4.1: Distribution of SRCC values for the surveyed NR-IQA methods.

striking upward trend in scores is observed. On these image sets, QualiCLIP, like TOPIQ, exhibits a clear tendency towards inverse ranking, assigning progressively increasing scores to images with progressively decreasing quality. The inverse rankings are less pronounced but still present on the rainy image sets. QualiCLIP achieved mean SRCCs of -0.121 and -0.115 on the rainy image sets from Otaniemi and Munkkivuori, respectively.

Finally, the assigned scores of IL-NIQE are shown in Figure 4.6. Unlike the other NR-IQA methods, IL-NIQE assigns lower scores to images it deems to have higher quality and higher scores to images it deems to have lower quality. Figure 4.6 demonstrates a clear increasing trend in scores across both locations and distortion types. This trend is more substantial for the foggy image sets, as reflected in the higher mean SRCC of 0.999. In contrast, the trend is less evident for the rainy image sets, which have a lower mean SRCC of 0.892.

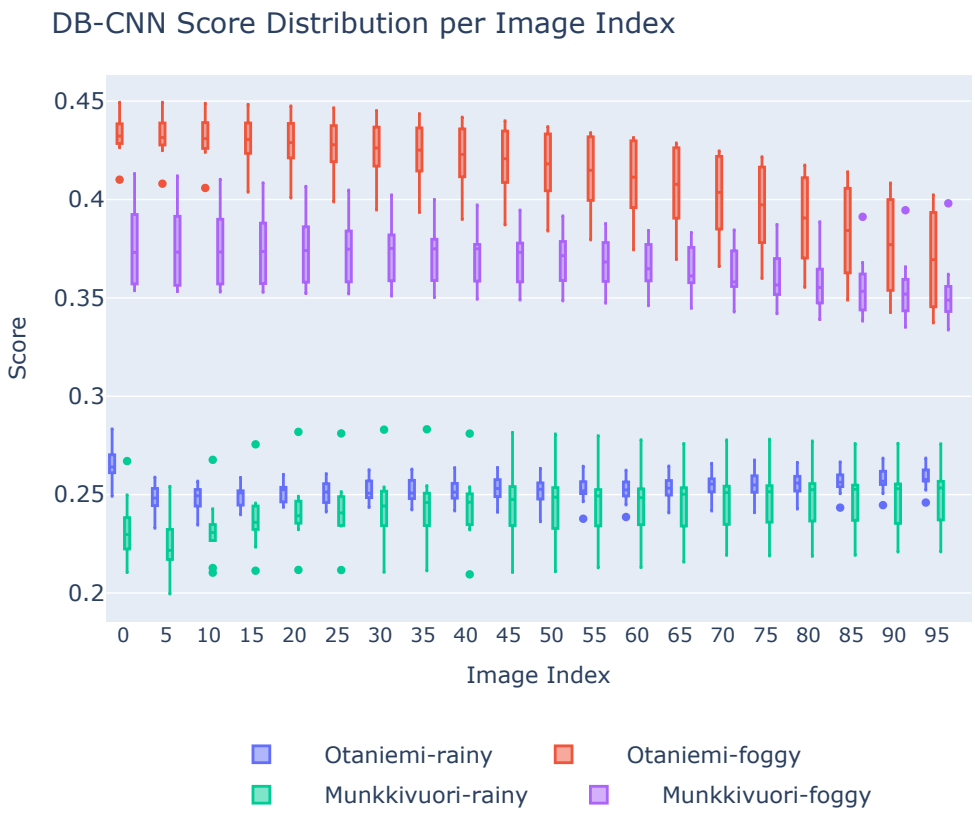


Figure 4.2: Boxplot of DBCNN scores for every fifth image index, showing score distributions across weather and location variations

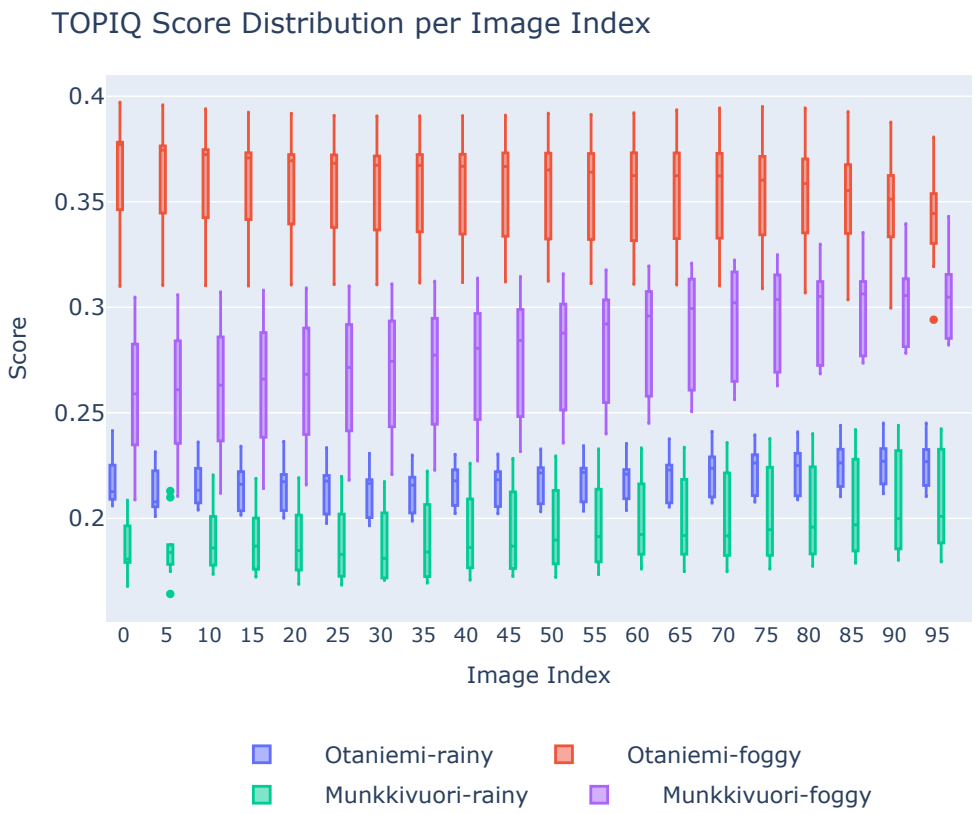


Figure 4.3: Boxplot of TOPIQ scores for every fifth image index, showing score distributions across weather and location variations

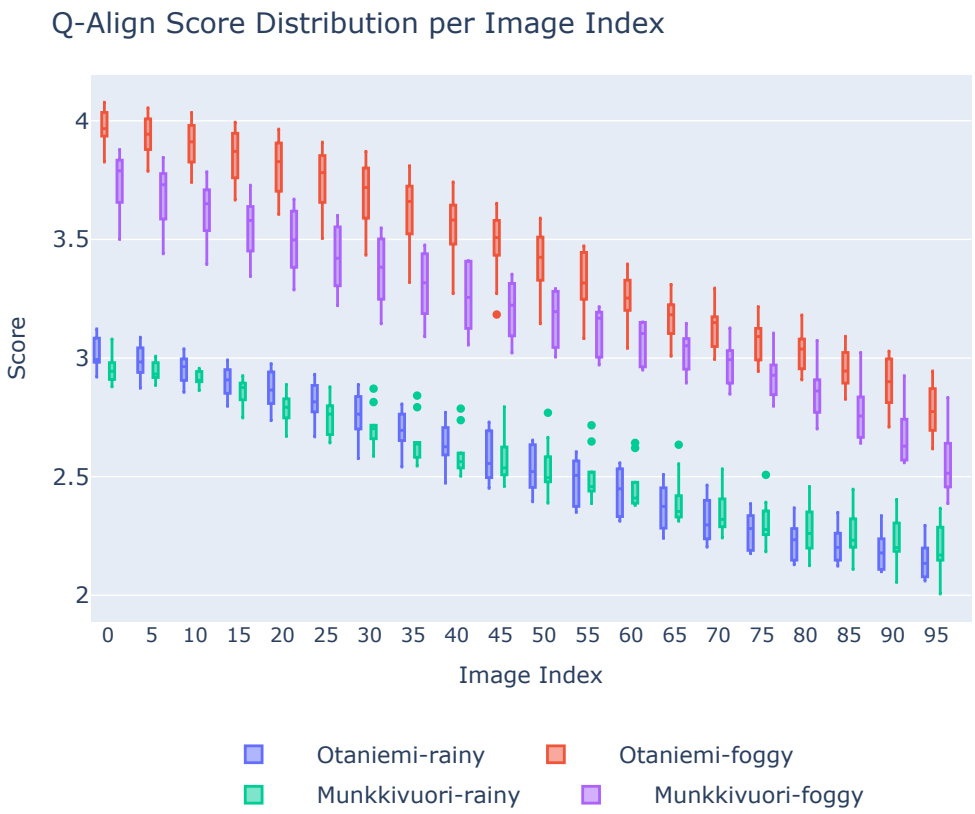


Figure 4.4: Boxplot of Q-Align scores for every fifth image index, showing score distributions across weather and location variations

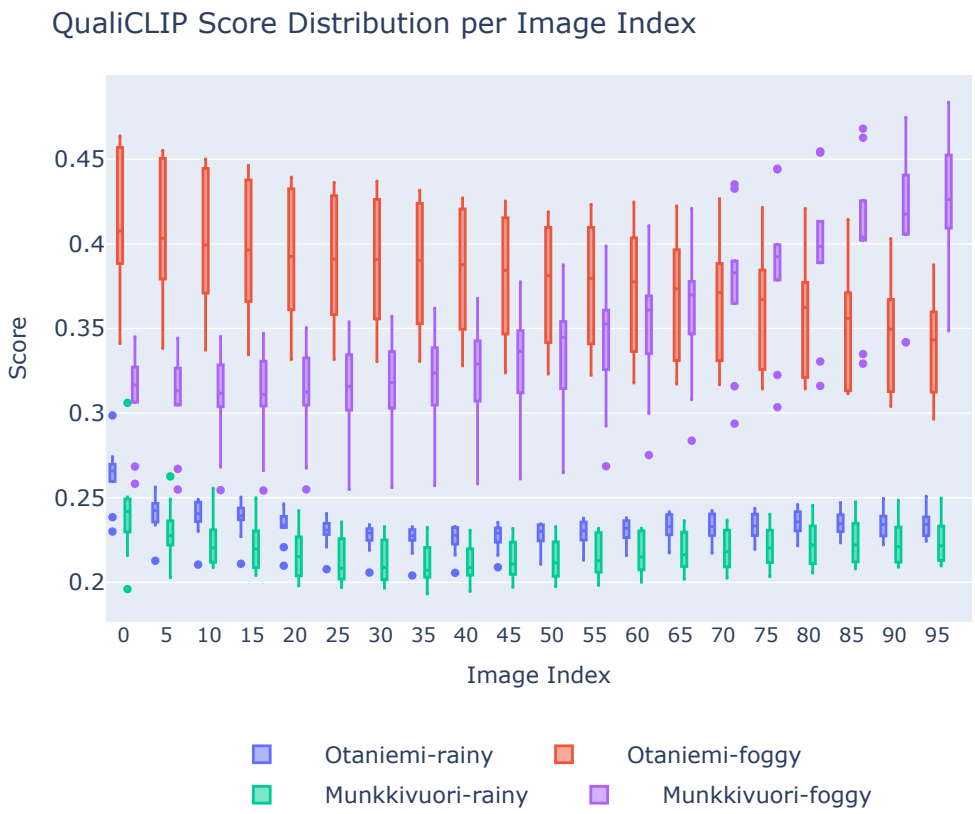


Figure 4.5: Boxplot of QualiCLIP scores for every fifth image index, showing score distributions across weather and location variations

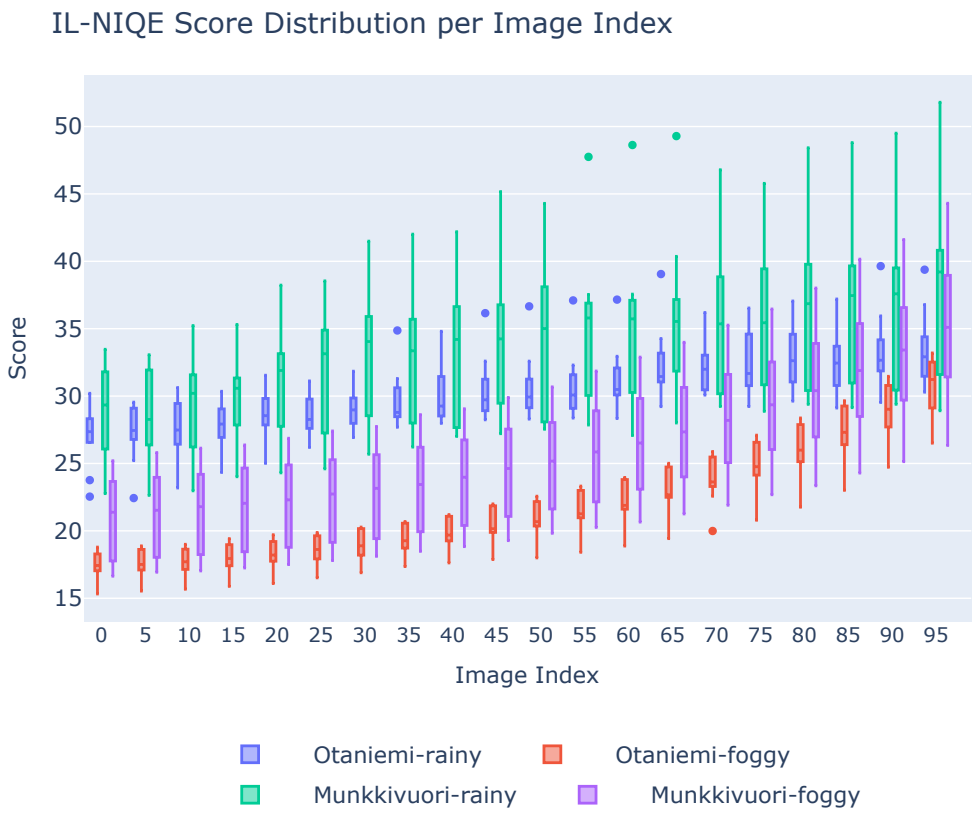


Figure 4.6: Boxplot of IL-NIQE scores for every fifth image index, showing score distributions across weather and location variations

4.1.1 Statistical tests

The result of the Friedman test is 109.11, with a p-value of 1.126×10^{-22}, meaning that the null hypothesis of no significant differences between the methods can be rejected at a 1% significance level. Thus, it can be demonstrated with strong statistical significance that at least one method consistently outperforms the others. Following the Friedman test, the results from the pairwise Wilcoxon signed-rank tests are found in Table 4.2. In this table, the model pair column lists each pairwise combination between the models, where the difference is computed as the first method minus the second. The statistic column lists the Wilcoxon signed-rank statistics, $\sum R^+$ lists the sums of the positive ranks, $\sum R^-$ lists the sums of the negative ranks, and the corrected p-value column lists the p-values after the Holm-Bonferroni correction.

The column of corrected p-values shows that all values, except for DBCNN - QualiCLIP, are below 0.01. This means that, at the 1% significance level, the null hypothesis of equal performance can be rejected for all model pairs except DBCNN and QualiCLIP. In the first four rows of Table 4.2, the negative rank sum $\sum R^-$ is substantially greater than the positive rank sum $\sum R^+$ for all four comparisons. Therefore, it can be concluded that TOPIQ is outperformed by all other models. The fifth row shows that there is no significant difference between DBCNN and QualiCLIP. The next two rows show that DBCNN is outperformed by both Q-Align and IL-NIQE. Similarly, the two rows that follow demonstrate that QualiCLIP is outperformed by both Q-Align and IL-NIQE. Finally, the last row indicates that Q-Align outperforms IL-NIQE. Based on the statistical tests, Q-Align achieved the best performance, followed by IL-NIQE. DBCNN and QualiCLIP were tied for third place, while TOPIQ performed the worst.

Model Pair	Statistic	$\sum R^+$	$\sum R^-$	Corrected p-value
TOPIQ − DBCNN	96	96	-724	0.000
TOPIQ − QualiCLIP	74	74	-746	0.000
TOPIQ − Q-Align	0	0	-820	0.000
TOPIQ − IL-NIQE	0	0	-820	0.000
DBCNN − QualiCLIP	374	445	-374	0.646
DBCNN − Q-Align	32	32	-788	0.000
DBCNN − IL-NIQE	39	39	-741	0.000
QualiCLIP − Q-Align	4	4	-816	0.000
QualiCLIP − IL-NIQE	6	6	-814	0.000
Q-Align − IL-NIQE	96	724	-96	0.000

Table 4.2: Results of the Wilcoxon signed-rank tests after Holm-Bonferroni corrections.

4.2 NR-PCQA results

The mean SRCC and KRCC values of MM-PCQA and MS-PCQE are found in Table 4.3. The standard deviations of the various methods are found in the parentheses. Furthermore, each p-value is from a two-tailed permutation test evaluating whether the mean differs significantly from 0. All the SRCC and KRCC values in Table 4.3 differ from 0 at the 1% significance level. Moreover, MM-PCQA has a mean SRCC value of 0.172 with a standard deviation of 0.401. This suggests a small and positive correlation between its assigned scores and the ground-truth qualities of the point clouds. Additionally, a standard deviation of 0.401 indicates that MM-PCQA exhibits inconsistent performance across the 40 point cloud sets. MS-PCQE exhibits a stronger negative correlation with the ground-truth qualities. It has a mean SRCC value of -0.294 with a standard deviation of 0.343. This shows that MS-PCQE, similar to some NR-IQA methods mentioned earlier, consistently rates lower-quality point clouds as having higher quality than those of better quality. However, MS-PCQE is also inconsistent across the image set, with a standard deviation of 0.343.

Table 4.3 shows the mean SRCC and KRCC of MM-PCQA and MS-PCQE on the fog-distorted point clouds. MM-PCQA yielded a very weak positive correlation, with a mean SRCC of 0.172 and a mean KRCC value of 0.129. However, the high standard deviations of 0.401 and 0.312 suggest inconsistent values across point cloud sets. MS-PCQE achieved a slight negative correlation, with a mean SRCC of -0.294 and a mean KRCC of

Method	SRCC (SD)	KRCC (SD)
MM-PCQA	0.172** (0.401)	0.129** (0.312)
MS-PCQE	-0.294** (0.343)	-0.220** (0.275)

* $p < 0.05$ ** $p < 0.001$

Table 4.3: Mean SRCC and KRCC values of the NR-PCQA techniques on the set of fog-distorted point clouds. The p-values are obtained from permutation tests evaluating whether the mean SRCC and KRCC differ significantly from 0.

-0.220. Similar to MM-PCQA, MS-PCQE exhibited high variability with standard deviations of 0.343 and 0.275, respectively.

Figure 4.7 plots the distribution of SRCC scores for both MM-PCQA and MS-PCQE across weather distortions and time of day. MM-PCQA generally achieved higher SRCC scores on point cloud sets recorded in clear weather and during daytime. The mean SRCC for point clouds captured in clear weather was 0.316, while for those captured during daytime, it was 0.237. However, both cases had high variability, with standard deviations of 0.325 and 0.395, respectively. MM-PCQA exhibited near-random performance on point clouds captured in rainy and nighttime conditions, with mean SRCC values of just 0.028 and 0.051, respectively. As with clear weather and daytime conditions, there was also high variability, with the standard deviation of 0.418 for rainy sets and 0.384 for nighttime sets.

In contrast, MS-PCQE demonstrates greater consistency in its quality predictions across varying weather conditions and between daytime and nighttime settings. On point clouds captured in clear weather, MS-PCQE achieved a mean SRCC of -0.298 (STD = 0.321), while in rainy weather, it achieved a mean SRCC of -0.290 (STD = 0.364). Similarly, the point clouds captured in daytime yielded a mean SRCC of -0.293 (STD = 0.337), and for point clouds captured in nighttime, it achieved -0.297 (0.354). These results suggest that MS-PCQE performs equally poorly across different weather conditions and between day and night scenarios.

Figures 4.8 and 4.9 show boxplots of the scores of the NR-PCQA methods across weather and time of day. Both plots contain 10 x-values, corresponding to the 10 distortion severity levels within each point cloud set. Figures 4.8 and 4.8 show no clear upward or downward trend for MM-PCQA or MS-PCQE.

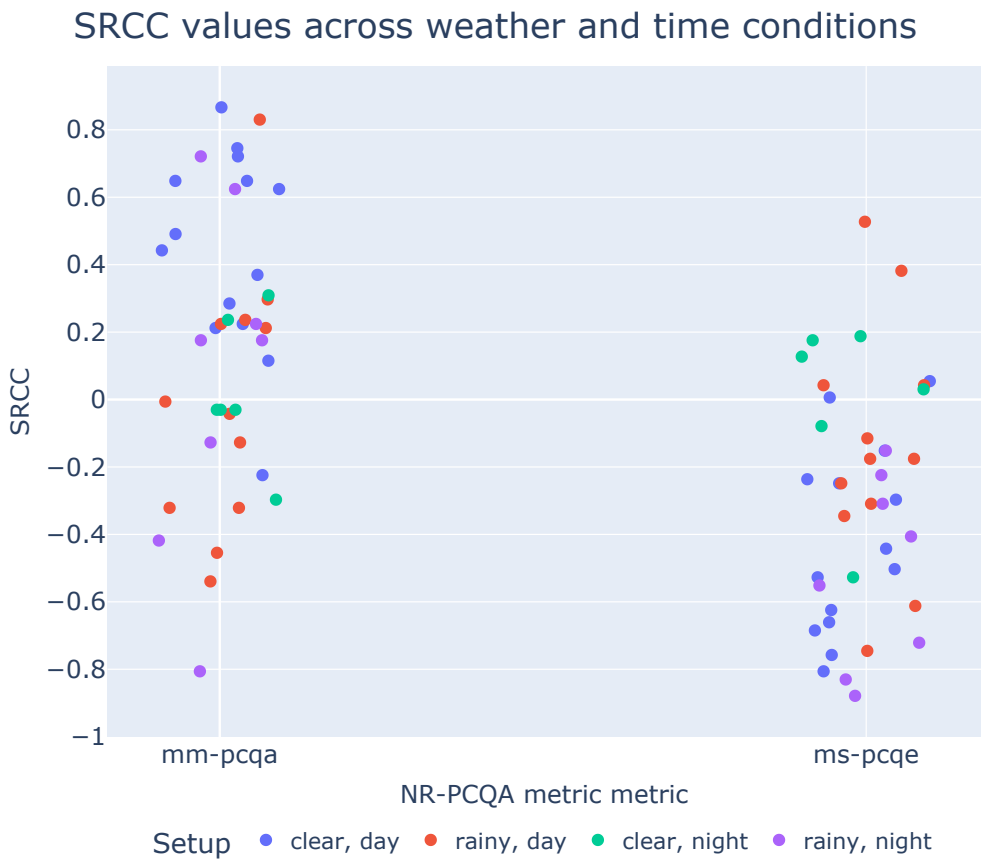


Figure 4.7: Distribution of SRCC values for the surveyed NR-PCQA methods.

4.2.1 Statistical tests

A two-sided Wilcoxon signed-rank test is performed to determine whether MM-PCQA and MS-PCQE differ significantly in performance. For this test, the SRCC values of each method across the 40 image sets are used. The resulting statistic is 90.5 with a p-value of $1.750 * 10^{-5}$, meaning that the null hypothesis that the methods perform equally well can be rejected at the 1% significance level. In the test, the SRCC values of MM-PCQA are subtracted from those of MS-PCQE. The resulting positive rank sum $\sum R^{+}$ is 90.5 while the negative rank sum $\sum R^{-}$ is -729.5, indicating that MM-PCQA significantly outperforms MS-PCQE.

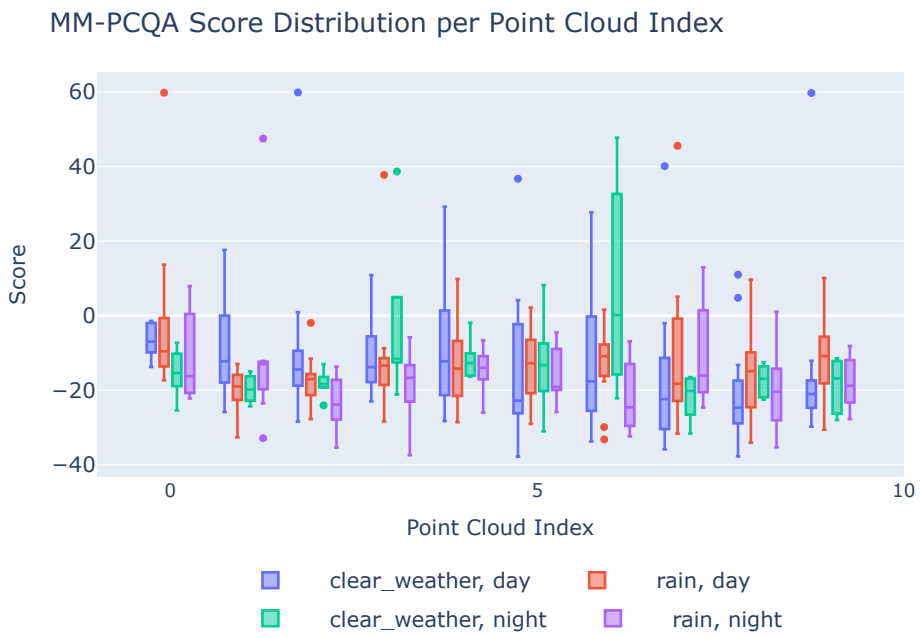


Figure 4.8: MM-PCQA score distributions across distortion levels and conditions.

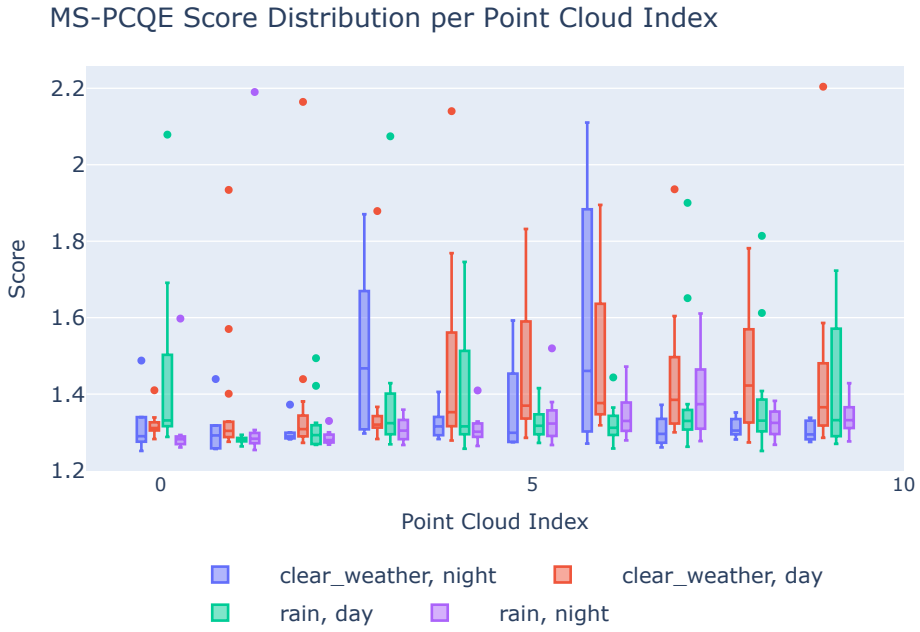


Figure 4.9: MS-PCQE score distributions across distortion levels and conditions,

Chapter 5

Discussion

Regarding the NR-IQA results, it is not surprising that Q-Align outperforms the four other evaluated methods. As a state-of-the-art, large multimodality model, Q-Align has demonstrated very strong performance on conventional NR-IQA benchmarks [36]. Q-Align was also trained on twelve NR-IQA datasets that contained both authentic and synthetic distortions. While training on both authentic and synthetic datasets typically reduces the accuracy of an NR-IQA model [36], the authors of Q-Align report that their model avoids this issue. This exposure to such a wide range of image content likely contributed to the superior performance of Q-Align in the study.

What is more surprising is the poor performance of another state-of-the-art NR-IQA method, namely QualiCLIP. As mentioned earlier, QualiCLIP exhibited slight inverse performance on the rainy image sets, near-perfect performance on the foggy Otaniemi image sets, and near-perfect inverse performance on the foggy Munkkivuori image sets. Overall, the lack of performance is surprising given that QualiCLIP has achieved impressive results on traditional NR-IQA benchmark datasets [35]. Additionally, the training process of QualiCLIP much resembles the evaluation process used in this thesis. QualiCLIP was trained by distorting images with increasing levels of noise and learning by ranking the distorted images. This is quite similar to the evaluation in this thesis. One possible reason why QualiCLIP performed poorly is that the degradations used in this study differ from those used to train QualiCLIP. QualiCLIP was trained using 24 distortion types, none of which included weather-related noise. This could have had a negative impact on performance.

However, the fact that QualiCLIP was not trained on weather-related distortions cannot explain the significant performance difference between the

foggy image sets across the two locations. On the foggy Otaniemi sets, QualiCLIP achieves a mean SRCC of 0.966, while on the foggy Munkkivuori sets, it achieves a mean SRCC of -0.977. This stark performance difference also occurs for TOPIQ, which achieves a mean SRCC of 0.576 on the foggy Otaniemi sets and a mean SRCC of -0.928 on the foggy Munkkivuori image sets. The markedly negative SRCC values of QualiCLIP and TOPIQ mean that they consistently assign higher scores to images with worse perceptual quality. An example of this can be seen in Figure 5.1. The discrepancy in performance across the two locations is likely attributed to differences in scenes and lighting. The images from Otaniemi are generally brighter and mainly consist of urban settings, while the Munkkivuori images are captured in rural environments and are darker. However, why these scene and lighting differences result in such dramatic changes in performance is unclear. Likewise, it is not clear why TOPIQ and QualiCLIP consistently assign higher scores to poor-quality images.

Another notable result is the consistent performance of IL-NIQE. Despite relying on NSS, a somewhat antiquated technique in the field of NR-IQA, IL-NIQE outperforms QualiCLIP, DBCNN, and TOPIQ. This is noteworthy as QualiCLIP and TOPIQ, both clearly outperform IL-NIQE on conventional NR-IQA benchmarks [33, 35]. One possible reason for IL-NIQE's strong performance is that synthetic rain and fog introduce consistent and predictable deviations in luminance and contrast compared to natural high-quality images. IL-NIQE may effectively detect these deviations by computing MSCN coefficients, which capture local variations in luminance and contrast.

A critical methodological decision in the thesis was to avoid using human subjective scores as the ground truth when evaluating images and point clouds. This decision was necessary, as there are no existing datasets of images and point clouds in the AV domain with subjective scores. However, it also has the consequence that it is not possible to evaluate whether a method's score aligns with human scores. This is a limitation since NR-IQA and NR-PCQA methods are typically also assessed based on how closely their scores align with human scores. It is therefore possible that one of the surveyed NR-IQA methods performs well in ranking images or point clouds, but that its absolute scores do not align with human perception.

The choice to use synthetic rather than authentic weather distortions also had its advantages and disadvantages. The main advantage was that it enabled distorting images with precise severity levels, which was necessary to establish ground truth rankings of the image and point cloud sets. However, the main disadvantage of this approach is that real-world weather phenomena are often

(a) QualiCLIP score of 0.3047 and Topiq score of 0.2247

(b) QualiCLIP score of 0.4407 and Topiq score of 0.2861

Figure 5.1: Example where QualiCLIP and Topiq assign a higher score to a perceptually worse-quality image

more complex than what can be modeled using synthetic noise. For instance, rainy conditions can lead to droplets on the camera lens, which cause total or near-total blockages. The artificial rain generation cannot account for this. Furthermore, foggy images often display a non-uniform distribution of fog. Areas of an image with close objects may appear clearer, whereas areas that depict distant backgrounds are more affected by fog. This is also something that the synthetic noise generation does not consider. The fact that synthetic noise does not fully model real-weather phenomena can be problematic, as a model can perform well on synthetic data but still struggle with data captured in authentic weather conditions. Therefore, it is crucial that the use of synthetic distortions is complemented with testing on authentic data to ensure that these systems perform reliably under real-world conditions.

Given the large number of NR-IQA methods in the literature, a selection of methods was necessary to keep the evaluation manageable. The goal with the selection was to cover a broad range of techniques that used different paradigms for NR-IQA. This meant that only one method per paradigm was included in the thesis. While this covers a broad range of techniques, it does not cover the nuances within each paradigm. Therefore, it is essential to recognize that alternative methods within these paradigms may yield significantly different results. Different results could possibly also have been obtained by using the same methods, but with different training datasets.

The poor results of the NR-PCQA models indicate that the field of NR-PCQA is not yet mature enough to accurately assess AV point clouds. MS-PCQE exhibited a negative correlation with the ground-truth rankings, indicating a failure to accurately rank point clouds of varying distortion levels. Furthermore, while MM-PCQA achieved a statistically significant positive correlation, its mean SRCC value of 0.172 is too low to indicate practical usefulness. Additionally, MM-PCQA assigns negative scores to several point clouds in the dataset. This is noteworthy, as it does not occur for any point cloud in the MM-PCQA training set. This result could be due to the point clouds in the dataset being significantly different from those used to train MM-PCQA. This further highlights the limited usefulness of MM-PCQA on AV point clouds.

The reason for the NR-PCQA model's inability to rank point clouds is most likely that NR-PCQA methods are primarily designed to assess the quality of highly detailed colored point clouds. This is indicative by the fact that the NR-PCQA benchmark datasets all contain highly rich and detailed point clouds. For instance, in the WPC dataset [75], which was used to train MM-PCQA, the mean number of points in each point cloud is 1359007. In contrast, the

number of points per point cloud in the AV point clouds used in the thesis is in the tens of thousands. The loss of color information in the AV point clouds is probably also a factor that explains the lack of performance.

Chapter 6

Conclusions and future work

6.1 Future work

One way to extend this work is to create a benchmark NR-IQA dataset of authentically distorted AV images with corresponding subjective scores. This would enable the evaluation of NR-IQA methods not only using rank-based metrics, such as SRCC and KRCC, but also accuracy-based metrics, including the Pearson linear correlation coefficient and root mean square error, which are standard in the NR-IQA field. Using naturally occurring and authentic distortions would also provide insight into whether the results in this thesis generalize to the real world.

Another interesting research area is investigating the feasibility of online NR-IQA, which involves evaluating image quality in real-time. This approach has been used by Zhang and Eskandarian [76] and was missed during the literature review of this study. Zhang and Eskandarian [76] explored online NR-IQA of autonomous vehicles by creating the detection quality index (DQI), which scores AV images based on saliency maps and the performance of object detection algorithms. Based on this method, they trained a neural network, SPA-NET, to predict the DQI. Future research in online NR-IQA could explore combining the DQI method with one or more of the NR-IQA methods evaluated in this study.

To enhance the performance of NR-PCQA methods, future research could utilize fused LiDAR and image data to generate colored point clouds. This process involves calibrating the LiDAR and camera, projecting LiDAR points onto the 2D image plane, and then recording the RGB values of each 3D point based on the corresponding pixel in the image. Given that both NR-PCQA methods evaluated in the thesis require colored point clouds, their

performance would likely be improved with authentic color information rather than simulated color information, as used in this thesis.

6.2 Conclusions

This thesis evaluated a set of NR-IQA and NR-PCQA methods on weather-distorted data used in the field of autonomous vehicles. The evaluation centered on synthetic weather distortions of images and point clouds, where the relative quality rankings were known. In the case of images, synthetic rain and synthetic fog were employed, whereas in the case of point clouds, only synthetic fog was utilized. Among the five evaluated NR-IQA methods, Q-Align and IL-NIQE demonstrated strong performance across both distortion types, suggesting that large multimodality models and NSS models are suitable for the task at hand. While both performed well, a Wilcoxon signed-rank test revealed that Q-Align ultimately outperformed IL-NIQE. The NR-PCQA evaluation showed that MM-PCQA outperformed MS-PCQE, achieving a weak, statistically significant correlation. However, this correlation is too weak to have any practical usefulness. Overall, the findings suggest that NR-IQA is more mature and reliable for applications in autonomous vehicles than NR-PCQA, which requires further research to be viable.

Chapter 7

Lessons learned

During the project, I learned the importance of keeping track of ideas, results, methods, and other thoughts in an organized manner. I was encouraged to write about my work in a blog throughout the thesis, and this turned out to be helpful in several ways. It helped me organize my thoughts and keep structured notes of the tools and methods I had used. It was also useful, when writing the thesis, to revisit my notes and see the reasoning behind the decisions I had made about tools, methods, etc. I also learned the importance of testing the viability of approaches and tools early on. The initial methods I explored for the thesis did not work, and identifying this early meant that I had time to pivot to an alternative approach. Likewise, several of the distortion tools that I had planned to use turned out to be unsuitable for the thesis. If I had not tested the viability of these tools and approaches early on, I would not have had time to switch to better-suited options. Another lesson was realizing how much the available data shapes what you can do. The lack of ground-truth labels in the datasets put strong constraints on the method, which in turn led to several limitations to the study. This was not something I had considered before I started writing the thesis.

References

[1] B. Crisafulli, R. Guesalaga, and R. Dimitriu, "Consumers' adoption of autonomous cars as a personal values-directed behavior," *Journal of Business Research*, vol. 189, p. 115106, 2025. [Page 1.]

[2] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixao, F. Mutz *et al.*, "Self-driving cars: A survey," *Expert systems with applications*, vol. 165, p. 113816, 2021. [Page 1.]

[3] 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. [Page 1.]

[4] M. Bijelic, T. Gruber, F. Mannan, F. Kraus, W. Ritter, K. Dietmayer, and F. Heide, "Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 682–11 692. [Page 1.]

[5] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," *arXiv preprint arXiv:1907.07484*, 2019. [Pages 1 and 6.]

[6] Y. Suh, Y. Chung, and Y. Park, "Ai training data management for reliable autonomous vehicles using hashgraph," *Applied Sciences*, vol. 15, no. 11, p. 6123, 2025. [Page 1.]

[7] "The ROADVIEW Project - Developing robust automated driving in extreme weather conditions," https://roadview-project.eu/, accessed: 2024-11-28. [Pages ix, 2, and 18.]

[8] C. Fu, C. Mertz, and J. M. Dolan, "Lidar and monocular camera fusion: On-road depth completion for autonomous driving," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 273–278. [Page 5.]

[9] Y. Li and J. Ibanez-Guzman, "Lidar for Autonomous Driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, Jul. 2020. doi: 10.1109/MSP.2020.2973615 ArXiv:2004.08467 [cs]. [Online]. Available: http://arxiv.org/abs/2004.08467 [Page 5.]

[10] S. Royo and M. Ballesta-Garcia, "An overview of lidar imaging systems for autonomous vehicles," *Applied sciences*, vol. 9, no. 19, p. 4093, 2019. [Page 5.]

[11] M. Liu, E. Yurtsever, J. Fossaert, X. Zhou, W. Zimmer, Y. Cui, B. L. Zagar, and A. C. Knoll, "A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook," *IEEE Transactions on Intelligent Vehicles*, 2024. [Page 6.]

[12] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 2023. doi: https://doi.org/10.1016/j.isprsjprs.2022.12.021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271622003367 [Page 6.]

[13] H. Gupta, O. Kotlyar, H. Andreasson, and A. J. Lilienthal, "Robust object detection in challenging weather conditions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 7523–7532. [Page 6.]

[14] A. Haider, M. Pigniczki, S. Koyama, M. H. Köhler, L. Haas, M. Fink, M. Schardt, K. Nagase, T. Zeh, A. Eryildirim *et al.*, "A methodology to model the rain and fog effect on the performance of automotive lidar sensors," *Sensors*, vol. 23, no. 15, p. 6891, 2023. [Page 6.]

[15] D. Kent, M. Alyaqoub, X. Lu, H. Khatounabadi, K. Sung, C. Scheller, A. Dalat, A. bin Thabit, R. Whitley, and H. Radha, "Msu-4s-the michigan state university four seasons dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 658–22 667. [Page 6.]

[16] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, 2020. doi: 10.3390/info11020125. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125 [Pages 6, 21, and 24.]

[17] A. B. Jung, "imgaug," https://github.com/aleju/imgaug, 2025, [Online; accessed 03-Jul-2025]. [Page 6.]

[18] S. S. Halder, J.-F. Lalonde, and R. d. Charette, "Physics-based rendering for improving robustness to rain," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 203–10 212. [Page 7.]

[19] 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. [Pages ix, 7, 27, 28, and 30.]

[20] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, pp. 973–992, 2018. [Page 7.]

[21] S. Sonawane and A. Deshpande, "Image quality assessment techniques: An overview," *International Journal of Engineering Research*, vol. 3, no. 4, 2014. [Page 7.]

[22] Y. ZHOU, Z. ZHANG, W. SUN, X. MIN, and G. ZHAI, "Perceptual quality assessment for point clouds: A survey," *ZTE Communications*, vol. 21, no. 4, p. 3, 2023. [Pages 7, 16, and 17.]

[23] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, pp. 1–52, 2020. [Pages 8 and 13.]

[24] Z. Wang and A. C. Bovik, "Modern image quality assessment," PhD Thesis, Springer, 2006. [Page 8.]

[25] Q. Mao, S. Liu, Q. Li, G. Jeon, H. Kim, and D. Camacho, "No-reference image quality assessment: Past, present, and future," *Expert Systems*, vol. 42, no. 3, p. e13842, 2025. [Pages 8, 13, and 17.]

[26] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012. [Page 8.]

[27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012. [Pages 8 and 9.]

[28] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3097–3100. [Page 8.]

[29] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015. [Pages 9, 30, and 33.]

[30] X. Yang, F. Li, and H. Liu, "A survey of dnn methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123 788–123 806, 2019. [Pages 9, 15, and 17.]

[31] Q. Mao, S. Liu, Q. Li, G. Jeon, H. Kim, and D. Camacho, "No-reference image quality assessment: Past, present, and future," *Expert Systems*, vol. 42, no. 3, p. e13842, 2025. [Page 9.]

[32] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020. doi: 10.1109/TCSVT.2018.2886771 [Pages 9, 30, and 33.]

[33] C. Chen, J. Mo, J. Hou, H. Wu, L. Liao, W. Sun, Q. Yan, and W. Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," *IEEE Transactions on Image Processing*, 2024. [Pages 10, 32, 33, and 52.]

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763. [Page 10.]

[35] L. Agnolucci, L. Galteri, and M. Bertini, "Quality-aware image-text alignment for real-world image quality assessment," *arXiv preprint arXiv:2403.11176*, 2024. [Pages 10, 20, 32, 33, 51, and 52.]

[36] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun *et al.*, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023. [Pages 10, 11, 15, 32, 33, and 51.]

[37] 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. [Pages 11, 13, 16, 17, and 32.]

[38] 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. [Pages 11 and 32.]

[39] 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. [Page 11.]

[40] Z. Zhang, W. Sun, X. Min, T. Wang, W. Lu, and G. Zhai, "No-reference quality assessment for 3d colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022. [Page 11.]

[41] Q. Liu, H. Yuan, H. Su, H. Liu, Y. Wang, H. Yang, and J. Hou, "Pqa-net: Deep no reference point cloud quality assessment via multi-view projection," *IEEE transactions on circuits and systems for video technology*, vol. 31, no. 12, pp. 4645–4660, 2021. [Page 12.]

[42] X. Chai and F. Shao, "Ms-pcqe: Efficient no-reference point cloud quality evaluation via multi-scale interaction module in immersive communications," *IEEE Transactions on Consumer Electronics*, 2024. [Pages 12 and 16.]

[43] Z. Zhang, W. Sun, X. Min, Q. Zhou, J. He, Q. Wang, and G. Zhai, "Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment," *arXiv preprint arXiv:2209.00244*, 2022. [Page 12.]

[44] 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. [Page 13.]

[45] S. Xu, S. Jiang, and W. Min, "No-reference/blind image quality assessment: a survey," *IETE Technical Review*, vol. 34, no. 3, pp. 223–245, 2017. [Pages 13 and 15.]

[46] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006. [Page 14.]

[47] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015. [Page 14.]

[48] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015. [Page 15.]

[49] Q. Yang, H. Chen, Z. Ma, Y. Xu, R. Tang, and J. Sun, "Predicting the perceptual quality of point cloud: A 3d-to-2d projection-based exploration," *IEEE transactions on multimedia*, vol. 23, pp. 3877–3891, 2020. [Page 15.]

[50] Q. Liu, H. Su, Z. Duanmu, W. Liu, and Z. Wang, "Perceptual quality assessment of colored 3d point clouds," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2022. doi: 10.1109/TVCG.2022.3167151 [Pages 16 and 33.]

[51] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, "Perceptual quality assessment of 3d point clouds," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3182–3186. [Pages 16 and 33.]

[52] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1040–1049. [Page 19.]

[53] I. Marsh, "Initial readiness assessment of specific datasets," https://roadview-project.eu/wp-content/uploads/sites/59/2024/05/ROADVIEW_Deliverable-4.5_v04.pdf, May 2024, online; accessed 16 April 2025. [Page 20.]

[54] Y. Poledna, M. F. Drechsler, V. Donzella, P. H. Chan, P. Duthon, and W. Huber, "Rehearse: adverse weather dataset for sensory noise models," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2451–2457. [Pages 20 and 21.]

[55] H. Mokayed, A. Nayebiastaneh, K. De, S. Sozos, O. Hagner, and B. Backe, "Nordic vehicle dataset (nvd): Performance of vehicle detectors using newly captured nvd from uav in different snowy weather conditions." in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5314–5322. [Page 21.]

[56] Y. Kim, B. Park, and I.-Y. Moon, "A study on object detection performance through data augmentation under adverse weather conditions," in *2024 International Conference on Sustainable Technology and Engineering (i-COSTE)*. IEEE, 2024, pp. 1–3. [Page 21.]

[57] A. Shiran, J. Li, Y. Liu, M. Hummel, O. Jenewein, and K. Bezboruah, "Water level detection in adverse weather conditions using security cameras," in *SoutheastCon 2024*. IEEE, 2024, pp. 187–193. [Page 21.]

[58] A. Ahar, S. Mahmoudpour, G. Van Wallendael, T. Paridaens, P. Lambert, and P. Schelkens, "A just noticeable difference subjective test for high dynamic range images," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6. [Pages 23 and 24.]

[59] 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. [Page 27.]

[60] 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. [Page 27.]

[61] R. H. Rasshofer, M. Spies, and H. Spies, "Influences of weather phenomena on automotive laser radar systems," *Advances in radio science*, vol. 9, pp. 49–60, 2011. [Page 27.]

[62] Z. Li, B. Xie, C. Chu, W. Li, and Z. Su, "No-reference geometry quality assessment for colorless point clouds via list-wise rank learning," *Computers & Graphics*, p. 104176, 2025. [Page 30.]

[63] GreenValley International, "Assign color to points – LiDAR360 mls documentation," 2025, accessed: 2025-06-24. [Online]. Available: https://www.greenvalleyintl.com/docs/lidar360mls/Main/PointCloudTools/PointCloud/AssignColorToPoints.html [Page 30.]

[64] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 1220–1230. [Page 32.]

[65] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5148–5157. [Page 32.]

[66] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563. [Page 32.]

[67] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 071–14 081. [Page 32.]

[68] C. Chen and J. Mo, "IQA-PyTorch: Pytorch toolbox for image quality assessment," [Online]. Available: https://github.com/chaofengc/IQA-PyTorch, 2022. [Page 32.]

[69] B. Wang, B. Liu, S. Liu, and F. Yang, "Vcisr: Blind single image super-resolution with video compression synthetic data," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4302–4312. [Page 32.]

[70] C. Wang, J. Pan, W. Lin, J. Dong, W. Wang, and X.-M. Wu, "Selfpromer: Self-prompt dehazing transformers with depth-consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 6, 2024, pp. 5327–5335. [Page 32.]

[71] M. Marozzi, "Some remarks about the number of permutations one should consider to perform a permutation test," *Statistica*, vol. 64, no. 1, pp. 193–201, 2004. [Page 34.]

[72] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006. [Page 34.]

[73] "Interpret all statistics for Friedman Test." [Online]. Available: https://support.minitab.com/en-us/minitab/help-and-how-to/statistics/nonparametrics/how-to/friedman-test/interpret-the-results/all-statistics/ [Page 34.]

[74] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures David J. Sheskin*, fifth edition ed. Chapman & Hall/CRC, 2020. [Page 35.]

[75] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, "Perceptual quality assessment of 3d point clouds," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3182–3186. [Page 54.]

[76] C. Zhang and A. Eskandarian, "A quality index metric and method for online self-assessment of autonomous vehicles sensory perception," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 13 801–13 812, 2023. [Page 56.]

TRITA – EECS-EX 2024:0000
Stockholm, Sweden 2025

www.kth.se

€€€€ For DIVA €€€€

{
”Author1”: { ”Last name”: ”Stenmark”,
”First name”: ”Victor”,
”Local User Id”: ”u100001”,
”E-mail”: ”vstenm@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
}
},
”Cycle”: ”2”,
”Course code”: ”DA231X”,
”Credits”: ”30.0”,
”Degree1”: {”Educational program”: ”Master’s Programme, Computer Science, 120 credits”
,”programcode”: ”TCSCM”
,”Degree”: ”Master’s degree”
,”subjectArea”: ”Technology”
},
”Title”: {
”Main title”: ”Evaluating NR-IQA and NR-PCQA Methods on Weather-Distorted Data in Autonomous Driving”,
”Language”: ”eng” },
”Alternative title”: {
”Main title”: ”Evaluering av NR-IQA och NR-PCQA metoder på data med väderinducerat brus inom autonoma fordon”,
”Language”: ”swe”
},
”Supervisor1”: { ”Last name”: ”Veronika”,
”First name”: ”Domova”,
”Local User Id”: ”u100003”,
”E-mail”: ”veronica.domova@gmail.com”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Supervisor2”: { ”Last name”: ”Marsh”,
”First name”: ”Ian”,
”E-mail”: ”ian.marsh@ri.se”,
”Other organisation”: ”RISE”
},
”Examiner1”: { ”Last name”: ”Li”,
”First name”: ”Haibo”,
”Local User Id”: ”u1d13i2c”,
”E-mail”: ”haiboli@kth.se”,
”organisation”: {”L1”: ”School of Electrical Engineering and Computer Science”,
”L2”: ”Computer Science” }
},
”Cooperation”: { ”Partner_name”: ”RISE”},
”National Subject Categories”: ”10201, 10206”,
”Other information”: {”Year”: ”2025”, ”Number of pages”: ”1,??”},
”Copyrightleft”: ”copyright”,
”Series”: { ”Title of series”: ”TRITA – EECS-EX” , ”No. in series”: ”2024:0000” },
”Opponents”: { ”Name”: ”A. B. Normal & A. X. E. Normalè”},
”Presentation”: { ”Date”: ”2022-03-15 13:00”
,”Language”:”eng”
,”Room”: ”via Zoom https://kth-se.zoom.us/j/ddddddddddd”
,”Address”: ”Isafjordsgatan 22 (Kistagången 16)”
,”City”: ”Stockholm” },
”Number of lang instances”: ”2”,
”Abstract[eng]”: €€€€

€€€€,
”Keywords[eng]”: €€€€
Autonomous Vehicles, No-Reference Image Quality Assessment, No-Reference Point Cloud Quality Assessment, Weather-distorted Images, Weather-distorted Point Clouds, Data Quality €€€€,
”Abstract[swe]”: €€€€
€€€€,
”Keywords[swe]”: €€€€
Autonoma fordon, No-Reference Image Quality Assessment, No-Reference Point Cloud Quality Assessment, Väderstörda bilder, Väderstörda punktmoln, Datakvalitet €€€€,
}

acronyms.tex

```
%%% Local Variables:
%%% mode: latex
%%% TeX-master: t
%%% End:
% The following command is used with glossaries-extra
\setabbreviationstyle[acronym]{long-short}
% The form of the entries in this file is \newacronym{label}{acronym}{phrase}
%                                      or \newacronym[options]{label}{acronym}{phrase}
% see "User Manual for glossaries.sty" for the  details about the options, one example is shown below
% note the specification of the long form plural in the line below
%\newacronym[longplural={Debugging Information Entities}]{DIE}{DIE}{Debugging Information Entity}
%
% The following example also uses options
%\newacronym[shortplural={OSes}, firstplural={operating systems (OSes)}]{OS}{OS}{operating system}

% note the use of a non-breaking dash in long text for the following acronym
%\newacronym{IQL}{IQL}{Independent -QLearning}

% example of putting in a trademark on first expansion
%\newacronym[first={NVIDIA OpenSHMEM Library (NVSHMEM\texttrademark)}]{NVSHMEM}{NVSHMEM}{NVIDIA OpenSHMEM Library}

%\newacronym{KTH}{KTH}{KTH Royal Institute of Technology}

%\newacronym{LAN}{LAN}{Local Area Network}
%\newacronym{VM}{VM}{virtual machine}
% note the use of a non-breaking dash in the following acronym
%\newacronym{WiFi}{-WiFi}{Wireless Fidelity}

%\newacronym{WLAN}{WLAN}{Wireless Local Area Network}
%\newacronym{UN}{UN}{United Nations}
%\newacronym{SDG}{SDG}{Sustainable Development Goal}

\newacronym{IQA}{IQA}{Image quality assessment}
```