Impact Of Windshields On Dense Perception Tasks

Master Thesis

BY

Prasannavenkatesh Balaji
Matriculation Number - 10043034

WORK DONE AT

Glass Laboratory and Advanced System Solutions

Volkswagen AG, Wolfsburg

Submitted On 31.05.2024

EXAMINERS

1. Prof. Dr.-Ing. Udo Nackenhorst (Leibniz Universität Hannover)
2. Prof. Dr.rer.nat. Alexander Braun (Hochschule Düsseldorf)

SUPERVISORS

1. Dominik Werner Wolf M.Sc. (Volkswagen AG)
2. Fynn Bensel M.Sc. (Leibniz Universität Hannover)

Plagiarism Declaration

I declare that this work is my own and I have not used any external material other than the provided sources with the appropriate citations. Furthermore, this work, or parts thereof, has not been previously submitted in the same or similar form to the examination office.

Hannover, 31.05.2024
.....
Place, Date

.....
Name, Family Name

Acknowledgements

I begin by thanking Dominik for his wonderful support throughout this thesis and the preceding internship. I thank him for his insights and especially the freedom he provided me that helped me explore lots of seemingly detached topics. I consider him a fundamentally strong researcher and my mentor now. I also thank Fynn who has supervised me from the university. He is someone who has been flexible and accomodative of all my constraints. I also thank Prof. Udo Nackenhorst and Prof. Alexander Braun for their guidance and agreeing to be examiners for my thesis.

I thank Boris, Hagen, Kai and Martin from the GLASS Lab for being cheerful and making me feel as part of the team. I thank Lars for being a really warm and approachable boss. I thank Steffen and Juan who have been a support to me and have offered me guidance whenever needed.

I thank all the students with whom I have worked with in the team for their friendship, especially Lyons for being a kind friend who has kept in touch with me over time and continues to offer me guidance. I end by thanking my friend and roommate Rohan whom I consider as one of the finest guys I have come across and of course my parents without whom none of this would have been possible.

Abstract

Windshields are present in every vehicle and an understanding of how the windshield impacts the images captured by a camera that is mounted behind a windshield is critical to the development of Level-4 ADAS vehicles. This work focuses on the impact of windshields on the performance of AI models. While seamless performance is essential in the real world, the trustworthiness behind the adoption of AI in a safety critical application like autonomous driving is highly dependent on the calibration of a model's predictions.

An understanding of the confidence of models is also rooted in the understanding of the uncertainties. While post-hoc calibration approaches offer a computationally efficient way to minimize the miscalibration of models' outputs, different calibration measures behave differently and produce different calibration loss surfaces. All these aspects are studied in detail in this work. Additionally, a new optical quality indicator Information Capacity is studied in an attempt to substitute it in the place of existing metrics like the Modulation Transfer Function (MTF). The various studies are explained in detail with their corresponding results.

Keywords: Spatial Frequency Response · Modulation Transfer Function · Point Spread Function · Uncertainty Quantification · Model Calibration.

Contents

List of Figures

List of Tables

Glossary

ADAS Advanced Driver Assistance Systems. 2, 9, 17

CMOS Complementary Metal Oxide Semiconductor. 15

DNNs Deep Neural Networks. 2, 4, 5, 17, 21

ESF Edge Spread Function. 5

FoV Field of View. 3, 47

mECE Mean Expected Calibration Error. 2, 5

mIoU Mean Intersection over Union. 2, 5

MTF Modulation Transfer Function. 2, 3, 5

OOD Out-of-Distribution. 1, 2

PSF Point Spread Function. 10

SFR Spatial Frequency Response. 5

1 Introduction

In the past decade, research in autonomous driving has been consistently improving in multiple countries and automobile manufacturers are competing with each other to develop not just a functional but also a reliable self-driving car. An essential step for the homologation of autonomous cars is to convince regulatory authorities like the United Nations Economic Commission for Europe (UNECE) of the reliability of the multiple features of Advanced Driver Assistance Systems (ADAS). This is critical because on road, the performance of a driving model constantly impacts not just the occupants but the entire domain around the vehicle.

The Society of Automotive Engineers (SAE) has defined a hierarchy encapsulating different levels of autonomy and in its taxonomy for driving automation systems defines six levels in which level-4 stands for 'high driving automation' [1]. The various SAE-defined levels of autonomous driving are better illustrated in Figure 1.1. The difference between level-3 and level-4 is that in level-4, the system is equipped with two key abilities - one to be fully self-reliant and function autonomously under certain conditions (with respect to location and time) and two, the ability to intervene and make essential decisions in a situation that demands it [2].

It is already well understood that the quantity and especially quality of data precedes almost all other requirements in machine learning. In the context of perception, the data is generally some form of sensor-based measurement, like a camera capturing images with an imaging sensor. An extensive study of the behaviour of these sensors with informative metrics for a qualitative and quantitative analysis of the recorded data naturally gains importance as the reliance on sensors for the vehicles' functions steadily grows.

Figure 1.1: The 6 levels of ADAS, as defined by SAE [3].

For the driving model to be fully autonomous under certain conditions would mean that there are infinitely many possibilities for an out-of-distribution event (OOD) to occur within that time period. An example of an OOD event for a driving model

is when it encounters a never-seen-before stray animal. In this case, while a human driver can clearly discern that although it is a new being on the road it is still an animal, an AI model could dangerously classify the animal as a pole for example. The model has to be trained or at least, tested for such an event. This is where uncertainty quantification becomes a branch of study of paramount importance in the context of autonomous driving. A careful and thorough understanding of the model's behaviour and uncertainty (or confidence) under an OOD event is essential. OOD events are not the only events that pose a serious threat to the performance of an AI model in perception. A domain-shift or domain-drift could severely harm the functioning of an autonomous system. A natural domain-shift (or drift) could be defined as a subtle, almost imperceptible change to the input distribution arising from some physical phenomena like weather which have a systematic impact on the data that is being fed as input. A windshield mounted before the ADAS-camera also causes a shift in the domain of the input to the neural networks, which is discussed in detail in Section 5.1.

1.1 Motivation

The fundamental motivation behind this work is to understand the impact of a windshield - how a windshield impacts the imaging process of a camera and then impact the performance of deep neural networks (DNNs) trained for a perception task. While the need for an understanding of DNNs behaviour under the influence of optical wavefront aberrations induced by windshields is critical, not much can be ascertained from metrics such as the mean Intersection over Union and the mean Expected Calibration Error (mIoU and mECE), especially from the perspective of a windshield manufacturer. A reason behind this is that they are relevant only within the context of machine learning and a strong and reliable correlation between metrics that are used to characterize optical quality and metrics that are used to quantify the accuracy and confidence of a neural network is required, before the metrics could be convincingly used to classify whether a windshield is fit to be assembled on an autonomous vehicle during mass production. The Fraunhofer Institute for Cognitive Systems (IKS) classifies three critical requirements or steps that needs to be met by every autonomous driving system (Level-3 and higher) [4]. As indicated in the process flowchart Figure 1.2, the working of any autonomous driving system starts with the imaging and subsequently an in-depth analysis of the "scene" surrounding the vehicle. This scene is obtained in the form of real-time imaging by the cameras mounted on the car, either inside the vehicle behind the windshield or sometimes outside the vehicle (depending on the design choice and other factors). Cameras inside the car behind the windshield capture images, with their optical quality influenced by the optical aberrations induced by the windshield [5]. The float glass process, the bending process of glass and polyvinyl butyral (a thermoplastic resin) used in laminating glass layers are the main contributors to these aberrations.

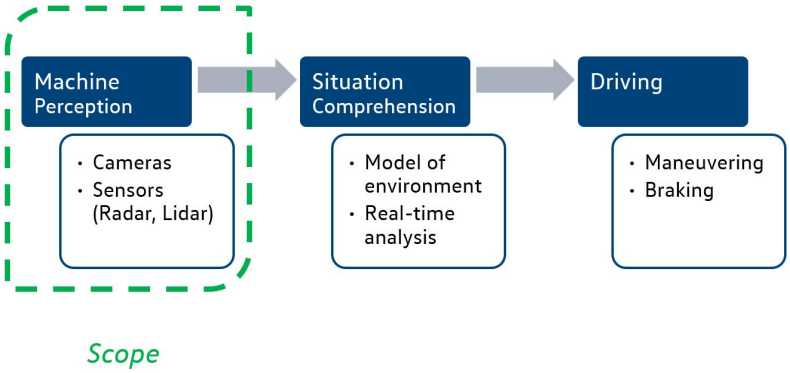

Figure 1.2: Three steps to self-driving

Thus, the windshield needs to be incorporated into the thorough analysis of how an image's quality is deteriorated. An important indicator of ADAS camera performance is the pixels-per-field-angle in the camera's field of view (FoV) [6]. As cars are designed to operate at high speeds, they are required to detect small objects at faraway distances well enough. This results in a resolution requirement. Modeling the intensity incident on every picture element in the sensor array with a Poisson distribution, it is found that the standard deviation of the spatial frequency response characterized by the MTF is roughly 2 % [7]. Therefore, uncertainties and systematic deviations higher than this magnitude need to be investigated and much smaller contributions can be neglected due to this physical precision limitation.

1.2 Objectives

This work started out as a continuation to the work of Wolf et al. [8] from Volkswagen's GLASS (Glass Laboratory and Advanced System Solutions) facility in Wolfsburg. This work was considered to be the base for my thesis. Also, the codebase for the optical threat model used to simulate optical wavefront aberrations and calculating the spatial frequency response from a slanted edge [9] were available in-house. With this, the open questions I attempted to answer through the course of my thesis are

1. To verify whether the optical threat model used by Wolf et al. [8] to simulate wavefront aberrations has the same impact on different datasets. Additionally the observation that the pure defocus dominates over aberrations from astigmatism induced by windshields needed to be verified. An extension of this was to attempt to prove the hypothesis that the relationship between optical quality metrics and model performance metrics could be independent of model architecture.

2. To explore the space of post-hoc calibration methods in machine learning. This was not part of the initial set of goals at the outset of this thesis. It was only

gradually understood how critical and computationally effective post-hoc calibration is, to improve the trustworthiness of a model's outputs.

3. To explore the space of other optical-quality metrics that could be superior to the metrics previously investigated and find if there are other metrics that establish an improved bijective relationship between optical quality and DNNs performance.

2 Literature Review

The first part of this chapter gives a review of existing literature on uncertainty quantification and model calibration approaches in the context of machine learning. The second part consists of a brief review of the literature on Information Capacity, most of which is from the extraordinary work of Claude Shannon and Imatest LLC, a company that develops image quality assessment software.

The ISO (International Organization for Standardization) set up a technical committee (TC-42) dedicated towards formulating standards, definitions and methodologies for electronic and still picture imaging. This committee, since its inception in 1947 has been publishing test charts which imaging equipment manufacturers have used to test, calibrate and benchmark their products. The committee, which met in 1991, proposed to develop a standard for measurement methodologies related to spatial resolution. This paved the way to the publication of ISO 12233 in 2000 for the first time [10]. In earlier times, to compare the performance of two lenses or cameras, limiting resolution metrics were used. This enabled one to quantify spatial resolution but limited to one value per direction - either in Cartesian space (vertical and horizontal) or in polar space (radial and tangential).

It was later found that these metrics were misleading and were not congruent with perceived sharpness in the target. This led to the widespread adoption of metrics based on the modulation transfer function (MTF) [11]. Later, Reichenbach et al. [12] found that a single slightly slanted edge could be used as a robust target to measure the MTF [12]. This is because the slight tilt in the edge gave rise to the possibility of oversampling the ESF which greatly reduced bias in the measurement. This was recommended to the ISO and subsequently the ISO committee also adopted slanted edge SFR (spatial frequency response) as the primary measurement methodology. It was found in [8] that the relationship between MTF and mIoU and mECE (metrics that quantify the performance of a DNNs) were not bijective, for it to be used as an encompassing metric that characterized the quality of a windshield. This led to the exploration of other metrics that are available - like the Shannon Information Capacity.

2.1 Uncertainty Quantification in ML

Recent work on uncertainty quantification and more recently uncertainty disentanglement suggest that the additive decomposition of the total uncertainty associated with the predictions of a probabilistic classifier like a neural network, has its own merits and demerits. Hüllermeier et al. [13] elaborates on this decomposition excellently. The total uncertainty of the prediction is commonly quantified with the help

of Shannon entropy which has its foundations in information theory [14] (discussed later in Section (4.4).

$$\begin{aligned} \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}) &= \int \mathcal{P}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x}) \cdot \mathcal{P}(\boldsymbol{w})\ \mathrm{d}\boldsymbol{w} \\ \mathcal{H}\left[\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\right] &= \int \mathcal{H}\left[\mathcal{P}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x})\right] \cdot \mathcal{P}(\boldsymbol{w})\ \mathrm{d}\boldsymbol{w} \\ &\quad + \int \mathcal{D}_{\mathrm{KL}}\left[\mathcal{P}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x})||\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\right] \cdot \mathcal{P}(\boldsymbol{w})\ \mathrm{d}\boldsymbol{w}. \end{aligned} \tag{2.1}$$

From the decomposition of the total predictive uncertainty denoted by $\mathcal{H}\left[\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})\right]$ into two integrals, it should be understood that the first term represents the expected conditional entropy. This conditional entropy indicates the aleatoric component of uncertainty and the source of this is the stochastic nature of the process that generates the input data. The second term is defined by the expected Kullback-Leibler divergence and indicates the epistemic component of uncertainty. The KL divergence quantifies the mutual information gain about the model weights $\boldsymbol{w}$ if the unconditional distribution $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ was already known. The decomposition in Equation (2.1) considers the weight distribution for a specific architecture thereby constraining the hypothesis space. There is a possibility that this hypothesis space may not include the true functional relationship that underlies the data generation. If this is the case, then model uncertainty must be considered in addition to the approximation uncertainty described by $\mathcal{P}(\boldsymbol{w})$. The Bregman decomposition [15] tackles this by introducing another (third) term into the additive decomposition of the total predictive uncertainty. Since analytically computing the true function behind the data-generation is not tractable, the KL divergence between the true function and the posterior distribution must be approximated to estimate the bias [16].

A thorough understanding of model uncertainty and calibration is essential in safety critical applications of machine learning like autonomous driving. Understanding model calibration well is one leap towards a holistic understanding of model uncertainty. Guo et al. [17] discovered that modern neural networks are poorly calibrated. Poor calibration translates to the model's outputs being overconfident. A model with perfectly calibrated outputs would mean that if the model predicts 100 outcomes with 80 % confidence, then 80 of those predictions would be correct. In the work of Guo et al. [17], model miscalibration (in the context of multiclass classification) was visualized and quantified with the help of reliability diagrams (introduced by Caruana et al. [18]) and the Expected Calibration Error metric (ECE) respectively [19]. Reliability diagrams are a powerful way to visualize the gap between binned accuracy and confidence of the predictions and a perfectly calibrated model would produce a reliability curve that coincides with the identity function. The ECE, first introduced by Naeini et al. [19] is a scalar single-valued metric that delivers a similar message as a reliability diagram - it quantifies the gap between the accuracy and confidence in a prediction. It was also shown by Guo et al. [17] that model cali-

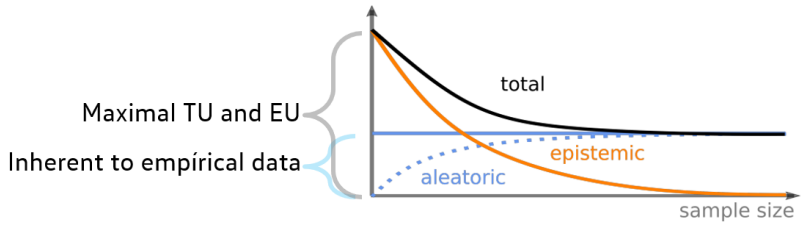

Figure 2.1: Illustration taken from the work of Wimmer et al. [20]. An important issue that challenges such an additive decomposition of the total uncertainty is that the aleatoric component should be constant in theory but the estimators for it show a dependency on the model's learning process. Hence, the epistemic contribution is overestimated by standard methods for small sample sizes.

bration is complex to understand and is dependent on multiple hyperparameters like the network depth and width, learning rate, weight decay etc and this led to a lot of following work on studying model calibration.

Approaches that target model calibration post-hoc are broadly classified into four groups (not necessarily exclusive). They are namely non-parametric, parametric, accuracy non-preserving and accuracy preserving calibration methods [21]. Generally, non-parametric methods alter the probability vectors' values thereby altering the model accuracy, which is undesirable in many studies which aim at estimating the true miscalibration of the model. Parametric approaches like Platt scaling [22] and temperature scaling [17] are relatively simple methods that use a global scalar referred to as the temperature T to rescale the logits vector produced by the final (pre-softmax) layer of the model. Temperature-scaling based methods, through logit-rescaling does not alter the ranking of the contents of the prediction vector $\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x})$, thus preserving the model accuracy [16]. This property can be described mathematically as

$$\underset{K}{\operatorname{argmax}}\ \sigma\left(\mathcal{L}\left(\boldsymbol{w}, \boldsymbol{x}\right)\right) \overset{acc}{=} \underset{K}{\operatorname{argmax}}\ \sigma\left(\frac{\mathcal{L}\left(\boldsymbol{w}, \boldsymbol{x}\right)}{T}\right) . \tag{2.2}$$

While recent studies have observed empirically that temperature scaling can be surpassed in performance by training-time approaches like label-smoothing [23], these are computationally intensive and also unfeasible to apply to black-box models [24]. These drawbacks make temperature scaling and its successor (in terms of improved expressiveness) parameterized temperature scaling a powerful candidate for model calibration and an important domain for ongoing research [21].

The seminal work of Nixon et al. [25] approached ECE critically and hinted at the fact that there might be a decoupling between minimizing the ECE and other calibration measures and exposed the pitfalls of using ECE as an estimator of the true calibration error. They attribute the binning of non-uniformly distributed data

into a fixed calibration range as one of the causes of this issue. They empirically observed the decoupling between the Negative Log-Likelihood (NLL) and the ECE as the source of different temperatures leading to the respective minimum values of these measures.

2.2 Theory of Communication

Claude Shannon introduced the theory of "information" applied to the field of communications in his seminal work published successively in 1948 and 1949 [26, 14]. The crux of this theory of information could be understood through the Shannon-Hartley theorem [14]. This theorem can be viewed as a typical application of the noisy-channel coding theorem (also referred to as Shannon's limit) to a continuous-time analog communications channel when Gaussian noise is present during transmission.

The theorem establishes that a determinable maximum rate of digital information transmission exists for a communication channel, for any degree of noise present to contaminate the desired signal. A rigorous proof to this theorem was first published by Feinstein et al. [27]. This led to the later understanding of how signals and transmission errors and different types of noise are correlated with one another. The intuition is that in the presence of noise that adversely impacts an image, the information content available in the image is affected negatively and this should be adequately reflected in the Information Capacity metric. Shannon et al. [26] define Information Capacity as the maximum capacity or rate at which information can be transmitted through a channel of an image at an arbitrarily low error rate and at a given level of noise present in the channel $c, q \in [0, 1]$.

3 Simulating Aberrations

Frequently evolving design choices of automotive manufacturers interested in developing vehicles with level-4 ADAS indicate that it is beneficial to bring the multitude of perception sensors (like the camera, LiDAR and Radar) together inside the car. This is advantageous since a lot of potential physical damage to the sensors due to inclement weather and accidents could be avoided. This would mean that the windshield is positioned in front of the camera and thus is part of the optical system. Modern cameras used in ADAS employ high resolution telephoto lenses and cover a narrow field-of-view. This increases the pixel resolution per field angle and in such a setting, a windshield has a greater impact on the imaging process.

Vehicle windshields typically possess a curvature and behave similar to how a thin lens would. Since the windshield manufacturing process is not as compliant to high-precision manufacturing tolerances as other lenses manufactured for microscopes/telescopes, it is natural to not expect a high degree of control on certain physical parameters of windshields, namely the curvature and thickness [28]. This leads to aberrations caused to the wavefront and at the same time, the high resolution demand of automotive imaging (going upto 4K horizontally) leads to an increased sensitivity to the adverse impact of windshields. This also means that the impact of windshields of varying types and qualities on the performance of ADAS cameras needs to be thoroughly analyzed. It is not convenient to only rely on physical experiments in facilities such as a light tunnel for such analyses. One needs to 'computationally' simulate the impact of windshields on cameras so that these analyses. This is what the optical threat models developed by Wolf et al. [8] through the course of his ongoing doctoral studies achieve. The equations in the following sections describing the theory follow a notation that is equivalent to a recent publication by Wolf et al. [8].

Fourier optics is an important field within the domain of optics that involves studying the Fourier spectrum (i.e. the frequency space) of the environment that is being imaged and how the spectral components of an object that is being imaged gets transmitted through the optical system. Fourier optics is fundamentally based on considering the wavefront as a superposition of plane waves. The reasoning behind the optical threat model used in simulating optical wavefront aberrations in this work lies in the theory of Fourier Optics.

3.1 Fourier Optics

Considering a field transmission function $f_t(x, y)$, an equivalent Fourier-domain description is expressed with the amplitude spectrum as a function of the spatial frequencies (u, v) as

$$F_t(u, v) = \frac{1}{(2\pi)^2} \iint f_t(x, y) e^{i2\pi ux + i2\pi vy} dxdy \tag{3.1}$$

A plane wave denoted by $\rho(\vec{x})$ having an amplitude U impinging on the object alters the distribution of the field behind the object which could be expressed as $U_o(x, y) = f_t(x, y) \cdot U(x, y)$ and this denotes that the information of the object has been "embedded" onto the wave. The Helmholtz equation which forms the crux of Fourier optics is defined as

$$\left(\Delta + k^2\right) \rho(\vec{x}) = 0 \quad \text{, with: } k := \frac{2\pi f}{c} = \frac{2\pi}{\lambda}, \tag{3.2}$$

where k denotes the angular wavenumber and $\rho(\vec{x})$ denotes an electromagnetic field wave. A spherical wave with an unit amplitude satisfying Equation (3.2) is commonly referred to as the free-space Green's function. Any optical system can be characterized by a Green's function $|h(\vec{x}_o)|^2$ such that the output of the system is modeled as the convolution of the input signal with $|h(\vec{x}_o)|^2$, for the incoherent light-incidence case. This Green's function describes the image of an infinitesimally narrow light pulse (a Dirac delta distribution) by the optical system and is commonly referred to as the point spread function function PSF.

$$\begin{aligned} \rho\left(\vec{x}_o\right) &= \iint_{\mathbb{R}^2} \left|h\left(\vec{x}_o - \vec{x}_s\right)\right|^2 \cdot \rho\left(\vec{x}_s\right) \mathrm{dx}_\mathrm{s}^2 \\ \Leftrightarrow \mathcal{F}[\rho](\vec{k}) &= \mathcal{F}\left[|h|^2\right](\vec{k}) \cdot \mathcal{F}[\rho](\vec{k}). \end{aligned} \tag{3.3}$$

Essentially, the resultant of any optical system is determined by the convolution of the raw captured image with this PSF. This convolution mechanism is the causal reason behind the validity of the superposition principle in wave-based treatment of optics. In Figure 3.1, a simple camera model with a thin lens is illustrated.

Assuming that the imaging system operates under incoherent light incidence and consequently there are no intereference effects present, the PSF $|h(\vec{x}_o)|^2$ of the system is expressed by the integral of the system's aperture function $P(\vec{x}_o)$ as

$$\begin{aligned} h\left(\vec{x}_o\right) &\approx \iint_{\mathbb{R}^2} P\left(\lambda d_z \vec{k}_{\tilde{a}}\right) \cdot \mathrm{e}^{-2\pi \mathrm{i} \vec{x}_o \cdot \vec{k}_{\tilde{a}}} \mathrm{dk}_{\tilde{a}}^2 \\ \Leftrightarrow h\left(\vec{x}_o\right) &= \mathcal{F}\left[P\left(\lambda d_z \vec{k}_{\tilde{a}}\right)\right] \text{, with: } \vec{k}_{\tilde{a}} := \frac{\vec{x}_a}{\lambda \cdot d_z}, \end{aligned} \tag{3.4}$$

where d_z quantifies the distance between the observation plane (at z_o) and the location of the aperture stop of the camera (at z_a).

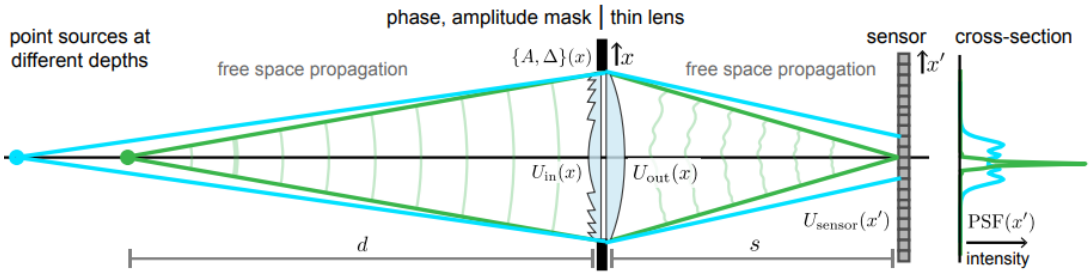

Figure 3.1: Camera model of a single convex thin lens with a focal length of f at a distance of s from the imaging sensor and capturing an object at a distance d behind the lens. Illustration from the work of Chang et al. [29].

With a design where the windshield is already a part of the imaging pipeline, an optical systems involved in autonomous driving is far from being modelled as being diffraction limited. Hence the aperture function from Equation (3.4) is extended to a generalized aperture function $\not{P}(\vec{x}_a)$ as a function of the wavefront aberration map as

$$\not{P}\left(\vec{x}_a\right) := P\left(\vec{x}_a\right) \cdot \exp\left[\frac{2\pi \mathrm{i}}{\lambda} \cdot W\left(\vec{x}_a\right)\right]. \tag{3.5}$$

In Equation (3.5), the term $W(\vec{x}_a)$ denotes the wavefront aberration map and this quantifies the difference in the optical path between the expected and the observed wavefront in the presence of a windshield in the optical path. Hence, one could rewrite both Equations (3.3, 3.4) in terms of the generalized aperture function as

$$\begin{aligned} \not{h}\left(\vec{x}_o\right) &= \mathcal{F}\left[\not{P}\left(\lambda d_z \vec{k}_{\tilde{a}}\right)\right] \\ \Rightarrow \mathcal{F}[\not{\rho}](\vec{k}) &= \mathcal{F}\left[|\not{h}|^2\right](\vec{k}) \cdot \mathcal{F}[\rho](\vec{k}). \end{aligned} \tag{3.6}$$

A large subset of optical systems' design, especially in ADAS applications employ cameras with a circular aperture stop of their objective lenses. Hence, the optical threat model developed by Wolf et al. [8] also assumes a circular pupil. Given that the generalized aperture function $\not{P}(\vec{x}_a)$ can be modeled directly as a function of the wavefront aberration map $W(\vec{x}_a)$, a parameterization of the aberration map with respect to a suitable coordinate system is essential to analyse how optical quality can be expressed as a function of the aberrations induced.

Zernike Decomposition: Zernike polynomials are chosen as the optimal choice of polynomials to provide a mathematical interpretation of what happens to an optical wavefront when it propagates through a system with a circular aperture. Zernike polynomials were first effectively used by Zernike et al. [30] in analysing circular mirrors. An optical wavefront is the surface of equivalent phase for radiation emanating from a monochromatic light source. Figure 3.2 illustrates an aberrated wavefront of a point source at a far away distance.

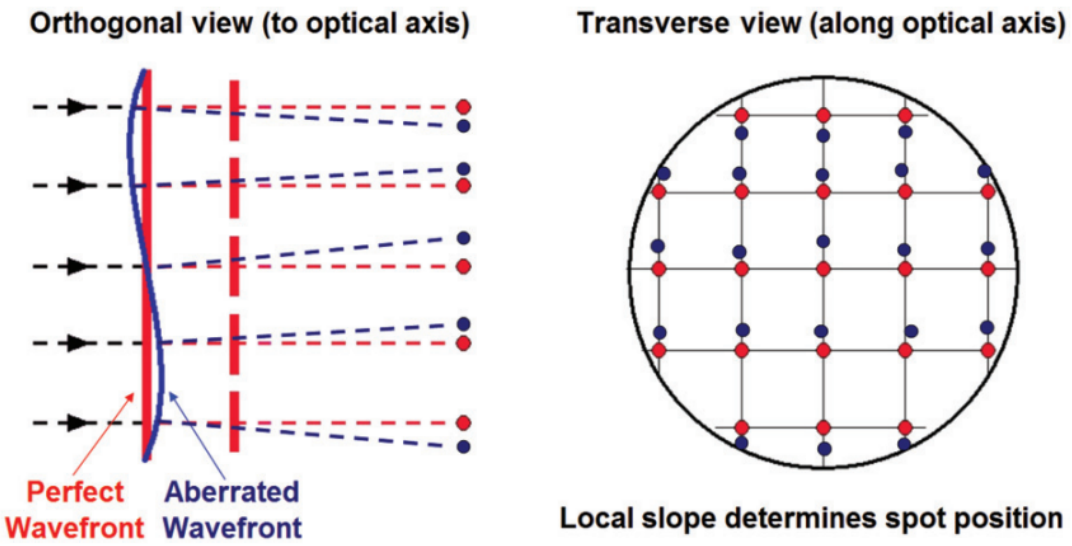

Figure 3.2: Orthogonal and transverse (to optical axis) view of the surface of equivalent phase (optical wavefront). Illustration from the work of Lakshminarayanan et al. [31].

The seminal work by Lakshminarayanan et al. [31] reviews this special set of orthogonal functions in great detail discussing the numerous benefits as well as the conditions under which Zernike polynomials offer an accurate mathematical description.

Advantageous Properties: Zernike polynomials, denoted by Z_n and parameterized by the polar coordinate system with the normalized radius ρ and polar angle ϕ offer certain benefits over other orthonormal polynomial families.

- Continuity: Continuous over the complete interior of an unit circle.
- Orthogonality: Orthogonal over the unit circle $\rho \in [0, 1]$.
- Independence: A direct result of using normalized Zernike polynomials is that each mode's coefficient can be equated to that mode's RMS contribution to the total wavefront error. The coefficients being independent of the number of polynomials already present in the expansion sequence means that infinitely many terms can be added to the expansion.

.

It is to be noted that in the work of Wolf et al. [8], the normalized version of the polynomials were used contrary to the standard definition in literature [31]. Using $Z_n(\rho, \phi)$ as the basis, a weighted decomposition of the aberration map W from Equation (3.5) can be performed as

$$W(\rho, \phi) = \sum_{n=0}^{\infty} \omega_n Z_n(\rho, \phi) \quad , \quad \omega_n := \langle W, Z_n \rangle \, . \tag{3.7}$$

Here n denotes the polynomials' index following the OSA/ANSI numbering scheme and the orthogonality criterion of terms of the expansion of W can be expressed in terms of the Kronecker delta by

$$\langle Z_b^a, Z_d^c \rangle := \int_0^{2\pi} \int_0^1 Z_b^a(\rho, \phi) \cdot Z_d^c(\rho, \phi) \cdot \rho \mathrm{d}\rho \mathrm{d}\phi = \delta_{p,q}. \tag{3.8}$$

with the indices of the Kronecker delta given by $p := \frac{b(b+2)+a}{2}$ and $q := \frac{d(d+2)+c}{2}$ respectively. With this, one could express the spatial frequency response i.e. the MTF as a function parameterized in terms of the Zernike polynomials' coefficients ω_n. The term corresponding to the zeroth order n = 0 denotes an offset of the wavefront in the longitudinal direction and the first order terms tip and tilt (n = 1 and n = 2) describe a deflection in the light beam. This results in a spatial displacement or shift of the entire image but cause no perturbations to the image's content at the pixel level. Hence the model developed by Wolf et al. [8] was restricted to consider only the polynomials corresponding to the second order n = 2. Table 3.1 tabulates the second order Zernike polynomials that fall within the low-order aberrations group. Figure 3.3 illustrates visually with color the various Zernike polynomials along with the commonly referred to optical effect it causes on images [31].

Radial Order (n)	Angular Frequency (m)	Zernike Polynomial (Z_n^m)
2	-2	$Z_2^{-2} = \sqrt{\frac{6}{\pi}}\rho^2 \sin 2\phi$
2	0	$Z_2^0 = \sqrt{\frac{3}{\pi}}\left(2\rho^2 - 1\right)$
2	$+2$	$Z_2^2 = \sqrt{\frac{6}{\pi}}\rho^2 \cos 2\phi$

Table 3.1: Zernike polynomials of the second order.

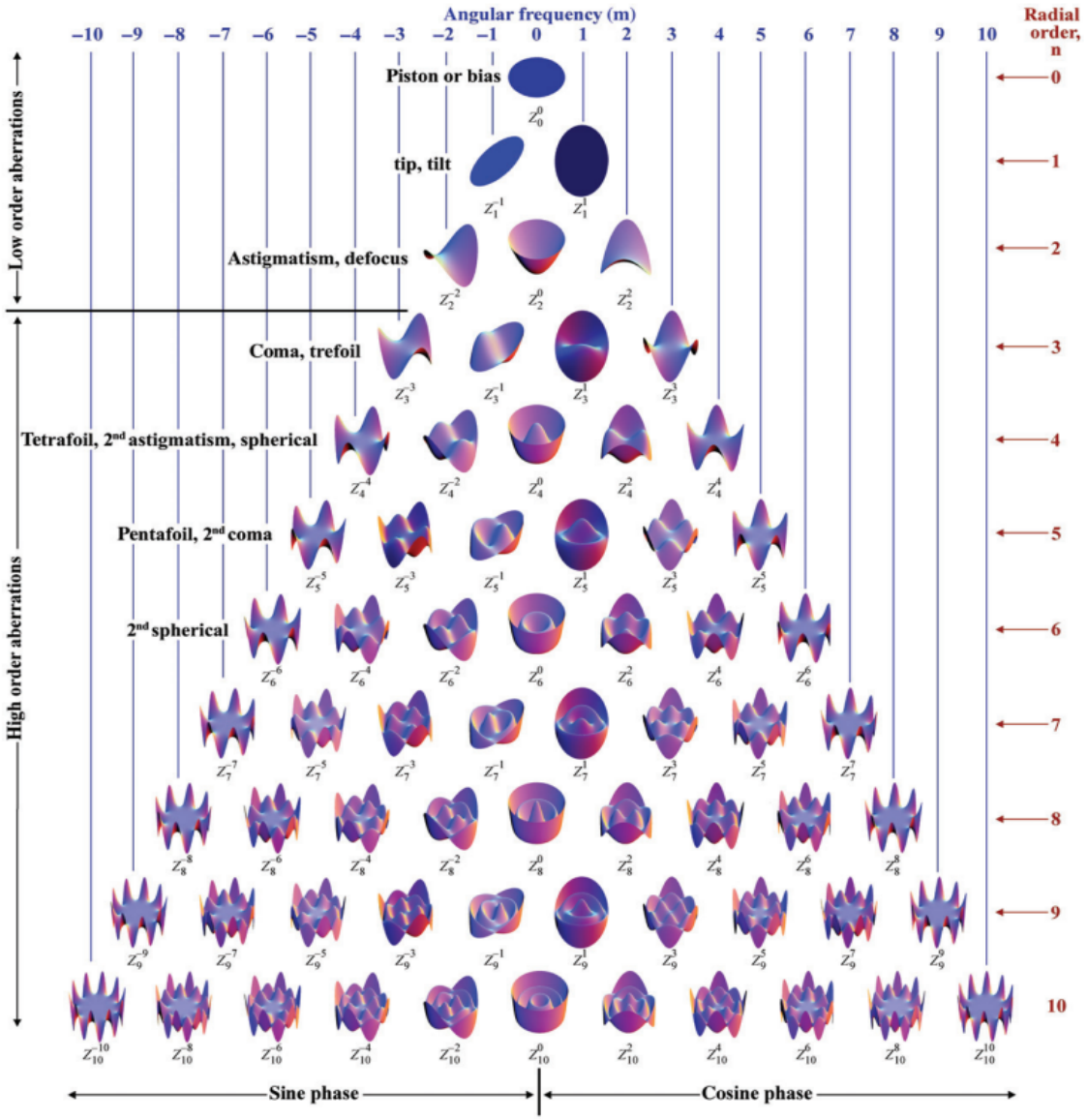

Figure 3.3: Zernike polynomials upto the 10^{th} order with the adverse optical effect it produces. Figure taken directly from the work of Lakshminarayanan et al. [31].

3.2 Implementation

This section briefly covers how the theory behind the threat model discussed in detail in Section 3.1 was implemented by Wolf et al. [8] in his doctoral studies to simulate optical aberrations on a single image.

1. The model assumes that the following data about the optical system is known apriori. The required parameters are:
   - The focal length and aperture stop of the camera used.
   - The pixel pitch τ and the horizontal and vertical dimensions (l_x and l_y) of the CMOS sensor of the camera.
   - The wavefront aberration map $W(\vec{x}_a)$, through measurement or optical simulation.
2. The domain of the MTF (frequency space) with the required discretization is given by
$$k_x, k_y \in \mathbb{R} \mid \left[\frac{-1}{2\cdot\tau}, \frac{-1}{2\cdot\tau}\right] \quad , \quad \Delta k_i = \frac{1}{\tau(N_i - 1)}. \tag{3.9}$$
3. The generalized aperture function $(\not{P})$ is built from $W(\vec{x}_a)$ and the available data using Equation (3.5).
4. From $\not{P}$, the perturbed PSF $|\not{h}|^2$ is calculated using Equation (3.3).
5. The spatial sampling rate of the observer and the aperture plane are calculated. The sampling rate for the PSF at the observer plane is determined by the pixel pitch $\Delta x_i^{(o)} = \tau$ and for $\not{P}$ the aperture plane discretization is determined by $\Delta \tilde{x}_i^{(a)} = \frac{\lambda \cdot d_z}{\tau \cdot (N_i - 1)}$ from Equation (3.3). The aperture array with the number of elements determined by the above sampling rate needs to be spanned for $W(\vec{x}_a)$ on the unit circle.
6. The perturbed image $\not{\mathcal{I}}$ is finally obtained as the convolution of the perturbed (non-diffraction limited) PSF over the true un-perturbed image $\mathcal{I}$.

Dataset	Focal Length [m]	f-Number	Sensor τ [m]	Sensor $l_y \times l_x$ [px]
KITTI	80×10^{-3}	8	4.65×10^{-6}	370×1224
Cityscapes	4.97×10^{-3}	1.6	2.2×10^{-6}	1024×2048
A2D2	5.49×10^{-3}	1.8	3×10^{-6}	1208×1920

Table 3.2: Optical threat model for different open-source driving-scene datasets.

In this work, three different datasets were used altogether. The optical system parameters of each dataset was substituted into the model to obtain three parameterizations of the optical threat model. Table 3.2 lists the optical system parameters of the datasets that were used.

3.3 Validation

A two-fold approach was performed by Wolf et al. [32], [8] to prove the physical validity and accuracy of the optical threat model developed. The Zernike coefficients which are assumed to be known apriori by the model were measured by means of a Shack Hartmann wavefront sensor on a test windshield.

- The MTF (Section 4.1.2) was measured with and without the windshield present in the optical path, based on the slanted edge method, as described in the ISO Standard ISO:12233 [9]. The MTF ratio (ratio of MTF with the windshield to the MTF without the windshield) was compared with the MTF that was produced by using the measured Zernike coefficients in the parameterization of the perturbed PSF $|\not{h}|^2$ with the developed threat model.
- Additionally, the optical power calculated from the aberration map parameterized by the threat model is compared with the results of a physical measurement of optical power on the test windshield using the Moire pattern technique [32]. The results were sufficiently close and comparable within the field of view of the camera and thus, the physical validity of the model was proven.

4 Metrics

This chapter describes the set of metrics describes the set of metrics that have been used through the course of this work in quantifying certain quantities of interest, in pursuit of studying the impact of optical wavefront aberrations. This chapter has been split into two. The first section covers the metrics that quantify optical quality and the second section covers the metrics that quantify the performance of DNNs both in terms of accuracy and in terms of calibration.

4.1 Optical Quality

Metrics that characterize optical quality are essential not just for a scientific study of the optical system and its design but also from a manufacturing standpoint. To re-emphasize, the goal behind the ongoing work at the GLASS laboratory facility at Volkswagen AG is to adopt an optical quality metric that reflects the true quality of the windshield and its impact on the various ADAS functions. Such a metric could then be conveniently used by both the vehicle manufacturer and the windshield manufacturer for homologation purposes. The metrics in this section (optical power, MTF and Strehl Ratio) are derived directly from the wavefront aberration map $W(\vec{x}_a)$ and the non-diffraction limited (perturbed) PSF $|\not{h}|^2$, as described in Equation (3.5) in Section 3.1.

4.1.1 Optical Power

In the context of windshields and its applications in the automotive industry, a flat plane wave is expected from the source and for windshields, optical power quantifies the spatial variation in the deflection vector. Refractive power is expressed as the second derivative of the wavefront aberration map with respect to the spatial directions and optical power $D_{x_i}(\vec{x}_a)$ is expressed as

$$\begin{aligned} D_{x_i}(\vec{x}_a) &= \frac{\partial^2}{\partial x_i^2} W(\vec{x}_a) \\ \Rightarrow \Delta W(\vec{x}_a) &= \sum_{i=1}^{d} D_{x_i}(\vec{x}_a) \\ \Rightarrow W(\vec{x}_a) &= \sum_{i=1}^{d} \Delta^{-1} D_{x_i}(\vec{x}_a) + \mathcal{C}(\vec{x}_a)\,. \end{aligned} \tag{4.1}$$

Here, the Laplacian operator is applied to the aberration map and the contribution of the aberration that fulfills the Laplace equation is impossible to retrieve. The optical power as a metric would be invariant to the Zernike polynomials that are harmonic functions [7]. This is a motivating factor behind the pursuit of a better metric that

could be used to formulate optical quality requirements for windshields.

4.1.2 Modulation Transfer Function

As discussed in detail in Section 3.1, the wavefront aberration map and the PSF (in the non-diffraction limited and incoherent light-incidence case) are used in describing the optical system. The MTF is a resolution metric derived directly from the perturbed PSF $\left|\not{h}\right|^2$

$$\not{\mathrm{MTF}}(\vec{k}) = \frac{\left|\mathcal{F}\left[|\not{h}|^2\right](\vec{k})\right|}{\left|\mathcal{F}\left[|\not{h}|^2\right](\vec{k}=0)\right|}, \tag{4.2}$$

where $\vec{k}$ denotes the spatial frequency (usually measured in this context in cycles per pixel). Here, the MTF is defined as the real part of the Fourier transform of the perturbed PSF and also normalized at zero spatial frequency ($\vec{k} = \vec{0}$). The reason behind this normalization is explained elaborately in the textbook by Goodman et al. [33].

4.1.3 Strehl Ratio

In an attempt to not depend on the MTF defined at only the half Nyquist frequency, the Strehl ratio (SR) takes into account the MTF across the entire spectrum from the origin to the Nyquist frequency in terms of the area enclosed by the function. Strehl ratio is defined as the ratio of the spectral integral of the MTF from the perturbed PSF and the diffraction-limited MTF as

$$\mathrm{SR}_{x_i} := \frac{\int_{\mathbb{R}} \not{\mathrm{MTF}}\left(k_{x_i}\right) \mathrm{dk_{x_i}}}{\int_{\mathbb{R}} \mathrm{MTF}\left(k_{x_i}\right) \mathrm{dk_{x_i}}} = \left.\frac{|\not{h}|^2}{|h|^2}\right|_{\vec{x}_o \overset{!}{=} \vec{0}}. \tag{4.3}$$

Strehl ratio can be equivalently defined also as the ratio of the aberrated PSF $|\not{h}|^2$ to the diffraction limited PSF (at the optical axis $\vec{x}_o \overset{!}{=} \vec{0}$).

4.1.4 Optical Informative Gain

The optical informative gain (OIG) introduced by Wolf et al. [8] aims to work as a function of the higher-order moments of the PSF, since the Strehl ratio in insufficient in providing information about the shape of the PSF. OIG is defined as

$$\mathrm{OIG}_{x_i} := \frac{\int_{\mathbb{R}} |\not{\mathrm{MTF}}|^2 \mathrm{dk_{x_i}}}{\int_{\mathbb{R}} |\mathrm{MTF}|^2 \mathrm{dk_{x_i}}} = \frac{\int_{\mathbb{R}} |\not{h}|^4 \mathrm{dx_{o_i}}}{\int_{\mathbb{R}} |h|^4 \mathrm{dx_{o_i}}}. \tag{4.4}$$

Equating the squared modulus of the MTF from the frequency domain to the fourth power of the PSF's modulus is a direct result of the proof of Plancherel's theorem

which states that the squared modulus of a function f is equivalent in the spatial (η) and the frequency (ξ) domain simultaneously:

$$\int_{-\infty}^{\infty} |f(\eta)|^2 d\eta = \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2 d\xi. \tag{4.5}$$

4.1.5 Information Capacity

This chapter contains a brief description of the theory behind calculating the information capacity metric from an image of a slanted edge target. Imatest LLC is a company that focuses on image quality measurements and benchmarking imaging equipment. They have been working and publishing methodologies to calculate information capacity from images of standard targets using different kinds of cameras [34]. In essence, there are two methods to calculate Information Capacity (IC).

1. Edge variance (EV) method - This method is based on calculating the spatially varying noise from the image itself. It is specifically recommended for bilaterally filtered images. Bilateral filtering involves sharpening the image (of a slanted edge target) near the edge and reducing the noise of the regions away from the edge. A clear peak in the noise amplitude $\sigma_s(x)$ is present in bilaterally filtered images. Commonly, consumer-usage DSLR cameras involve bilateral filtering.

2. Noise image (NI) method - The noise image method is based on calculating the noise from the binning procedure itself. This is recommended by Imatest LLC for uniformly processed images. An example of uniform processing is demosaicing the complete image based on the specific Bayer pattern of the imaging sensor. All images in my studies have been captured with a sensor with the R-C-C-B pattern [35].

   No visible peak is observed in the noise amplitude of uniformly processed images but the noise itself is proportional to the intensity of the pixel. Hence the noise is higher on the brighter side of the edge compared to the darker side. Although the NI method has been recommended for uniformly processed images used in machine vision applications, one needs to still calculate the spatially dependent noise from the EV method as it is used as the normalizing factor for the NPS in step 7 in Algorithm 1.

In the Noise Image method, four quantities of interest are calculated and consequently used to calculate the channel Information Capacity C. They are:

- 2-D noise image - The pixel-wise noise contents extracted from the raw noisy image. Like the raw image, the noise levels are also expressed in terms of intensity counts.

- 2-D noise spectrum - The squared modulus of the shifted fast Fourier transform of the noise image results in the 2-D noise spectrum, which is then radially averaged and normalized to obtain the 1-D noise spectrum denoted by NPS as a function of the spatial frequency.
- Noise Equivalent Quanta NEQ - NEQ has been used as a figure of merit in medical imaging applications previously, investigated in detail by Cunningham et al. [36]. Then it was recently revived and investigated in a completely different object (pedestrian) detection setting by Keelan et al. [37] who demonstrated that the NEQ is a better metric than the conventional Signal-to-Noise ratio (SNR) to evaluate spatial quality of digital imaging systems [37]. Given the assumption that noise present in an image is white noise (approximately flat power spectrum), the NEQ is defined by Keelan et al. [37] as $\frac{SNR^2}{pixel\ area}$ when the spatial frequency is expressed with the units cycles per pixel. Its units are therefore dimensionless.
- Finally, the Information capacity C_{NEQ} and the maximum Information Capacity C_{max} are calculated. It is also common to denote C_{NEQ} by C_4 in case the slanted edge target has a contrast ratio of 4:1.

An important distinction between Information Capacity and the previously discussed metrics in this section is that the Information Capacity metric is not possible to be calculated from a natural driving scene (an example of a natural scene is any image from a real-world driving scene dataset). Since it is directly dependent on the ISO:12233 algorithm for the MTF [9], there is an implicit requirement of slanted edges in the images and driving scenes may not have suitable slanted edges that could be used to calculate the MTF and subsequently the Information Capacity. While Zwanenberg et al. have devised a method to extract slanted edges from natural scenes and calculate the MTF [38], the datasets they use are different (not driving scenes) and are built in such a way that they contain multiple suitable slanted edges in every image. This is a current bottleneck with Information Capacity to be used as a metric alongside open-source datasets like KITTI/Cityscapes. The steps that have to be carried out to calculate C_{NEQ} and C_{max} have been elaborated in Algorithm 1 in Appendix B.

4.2 Windshield Quality

From each of the four metrics that have been defined in Section 4.1, a quality requirement can be formulated. This requirement is to be then set as a threshold to ensure quality in production. Tier-1 suppliers of glass require these requirements for ensuring their production quality and automotive manufacturers require them to decide on homologation standards that are based on thresholds set on these metrics. Equation (4.6) lists a set of quality requirements α, β, γ, δ, formulated as thresholds from

the metrics D, MTF, SR and OIG respectively with respect to the Cartesian coordinate system x, y, z.

$$\begin{aligned} \max\left\{|D_x|, |D_y|\right\} &\overset{!}{\leq} \alpha \\ \min\left\{\not{\mathrm{MTF}}_x, \not{\mathrm{MTF}}_y\right\}\big|_{k=\frac{Ny}{2}} &\overset{!}{\geq} \beta \\ \min\left\{\mathrm{SR}_x, \mathrm{SR}_y\right\} &\overset{!}{\geq} \gamma \\ \min\left\{\mathrm{OIG}_x, \mathrm{OIG}_y\right\} &\overset{!}{\geq} \delta. \end{aligned} \tag{4.6}$$

A comprehensive understanding of the behaviour of Information Capacity 4.1.5 and its relationship with windshield-induced optical aberrations and image noise has not yet been achieved. Hence, an optical quality threshold formulated using Information Capacity is not yet explored and hence not within the scope of this work.

4.3 Model Accuracy

DNNs (Deep neural networks) have been increasingly used in dense perception tasks like object detection and semantic segmentation in recent times, with recent developments revolving around the all-encompassing panoptic segmentation [39]. The accuracy reported by AI models is quantified with a number of metrics. Supervised training of models for semantic segmentation involves training the models with groundtruth labels annotated manually or with supervision-intensive annotation tools. Measuring the accuracy for a task like semantic segmentation where the accuracy is defined for every pixel in the image involves the pixel categorized as one of four groups: True Positives, True Negatives, False Positives and False Negatives. Involving these four groups, the metric Intersection over Union is defined as

$$\mathrm{mIoU} := \frac{1}{N_K} \sum_{i=0}^{N_K} \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \tag{4.7}$$

where N_K denotes the number of classes present in the analysis in total. Generally, the IoU is defined for each class present in the validation set's images and the mean over all the classwise IoUs is reported as the mean Intersection over Union mIoU over the batch. Generally the accuracy of a model during an inference study is indicated by the mIoU.

4.4 Uncertainty Measures

Mutliple measures could be used as an estimator of the predictive uncertainty. All these measures typically make use of the logits from the final layer of the model, transformed using the softmax function to have a distribution limited to the closed unit interval $[0, 1]$. Denoting the distribution of predictions for each pixel z over the

complete label space $\mathcal{Y}$ by $\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x})$, the following measures have been used in multiple works [40, 41].

- Variation Ratio - While it does not include the complete prediction vector into consideration, the variation ratio V is the simplest uncertainty measure obtained directly from the softmax outputs of a prediction. If the largest softmax output for a pixel z is interpreted as the confidence, then V denotes the complement of this confidence.

$$V = 1 - \max \hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x}) \ . \tag{4.8}$$

- Probability Margin - Probability margin M is defined as the complement of the difference between the first and the second largest softmax outputs for a pixel z . It is expressed as

$$M = 1 - V + \max_{y \in \mathcal{Y} \setminus \hat{\mathbf{p}}_1} \hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x}). \tag{4.9}$$

- Entropy - The Shannon entropy is a scalar measure calculated over a finite number of classes $\mathcal{Y} = y_1, y_2 ... y_K$ for every pixel z just using the softmax outputs. Shannon entropy is a commonly used measure for the predictive uncertainty due to certain advantageous properties: non-negativity, reaching a maximum for the case of a uniformly distributed $\mathcal{P}(\boldsymbol{y}|\boldsymbol{x})$ and invariance to permutations in the label space $\mathcal{Y}$ [42]. The Shannon entropy $\mathcal{H}$ in the discrete case is defined for each prediction instance as

$$\mathcal{H}\left[\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{x})\right] = -\frac{1}{\log\left(K\right)} \sum_{i=1}^{K} \hat{p}_i(y_i|\boldsymbol{x}) \cdot \log\left(\hat{p}_i(y_i|\boldsymbol{x})\right) \ . \tag{4.10}$$

  Equation (4.10) is normalized by $\log(K)$ which is the maximum possible value of Shannon entropy calculated over K classes. Maximum entropy corresponds to a prediction with the most uncertainty and that is achieved when the prediction is an uniform distribution over K. A brief proof of this normalization is provided in Section A.1 in Appendix A.

Figure 4.1 illustrates the difference between Shannon entropy $\mathcal{H}$ and Probability margin M.

4.5 Model Calibration

The metrics that have been employed in estimating the calibration quality (or conversely the miscalibration) of neural networks in this work are introduced in this section. The output from a probabilistic classifier $\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{x})$ on input data denoted by

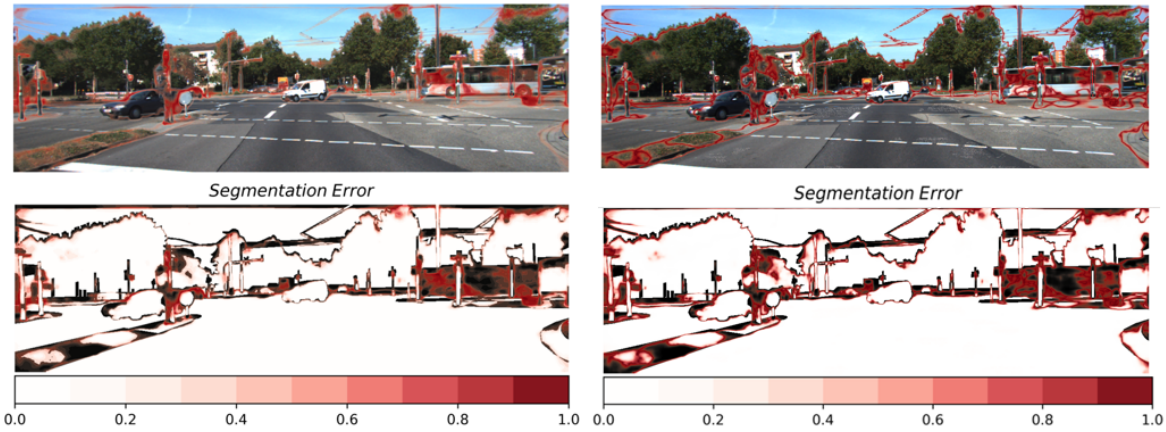

Figure 4.1: (L): The column on the left visualizes a heatmap of the Shannon entropy superimposed on the true image along with the segmentation error visualized as a binary mask. The white regions correspond to a right class prediction and the black regions correspond to a wrong prediction. (R): The column on the right illustrates a heatmap of the Probability Margin M. This metric places more emphasis on class boundaries since it is calculated as the difference between the top-2 softmax values [43]. This is noticeable in the image on the right with amplified uncertainty contours along class boundaries.

the vector $\boldsymbol{x}$ with its trained weights denoted by $\boldsymbol{w}$ is said to be 100% calibrated when the following state is achieved:

$$\mathbb{P}\left(\hat{y} = \operatorname{argmax} \hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}) \mid c = \max \hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x})\right) = c \;, \quad \forall c \in [0, 1] \;, \tag{4.11}$$

where $\hat{p} \in [0, 1]$ is the largest value in the probability mass vector, referred to as the softmax likelihood or softmax confidence. The class corresponding to the softmax likelihood is $\hat{y} \in \mathcal{Y}_K$ and c is the model accuracy [17] when the predictions are evaluated against the groundtruth. To illustrate this further, if an experiment is repeated 100 times identically with a softmax confidence $\max \hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x}) = 80\%$, then in 80 of those 100 experiments, the average accuracy of the predictions should be $c = 80\%$. This would mean that the softmax likelihood represents the true correctness of the model's predictions.

While this is one approach to quantify the miscalibration of model outputs in terms of the predictive accuracy, a complementary approach by Laves et al. [44] estimates the same miscalibration in terms of the predictive error. Over 100 experiments a model's predictions with an uncertainty $S\left(\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x})\right) = 20\%$ would lead to an average error of $q = 20\%$. This is described by rewriting Equation (4.11) in terms of uncertainty and error as

$$\mathbb{P}\left(\hat{y} \neq \mathbb{E}[\hat{y}] \mid S\left(\hat{\mathbf{p}}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{x})\right) = q\right) = q \;, \quad \forall q \in [0, 1] \;. \tag{4.12}$$

Here, S is the metric used to quantify the predictive uncertainty. In principle, any of the metrics discussed in Section 4.4 could be used as an estimate of the uncertainty associated with the prediction vector.

Since a single model is used in our studies and not an ensemble, the distribution over the weights $\mathcal{P}(\boldsymbol{\omega})$ is implicitly assumed to be given by a Dirac delta distribution $\delta(\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})$, where $\hat{\boldsymbol{\omega}}$ denotes the point-wise predictions for the weights of the trained network. An ensemble-based approach would result in a distribution obtained for approximating the weights from the independently trained ensemble members.

4.5.1 Expected Calibration Error

The difference between the model's accuracy and the softmax likelihood, referred to as the miscalibration is quantified by the expected calibration error (ECE), first introduced by Naeini et al. [45]. The ECE is defined as

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}\left(B_m\right) - \mathrm{conf}\left(B_m\right)\right| \ , \tag{4.13}$$

where the predictions from the final softmax function σ are binned into M bins B_m of equal width in the range $[0, 1]$ and the total number of predictions is denoted by n. The absolute difference between the average accuracy $\mathrm{acc}\left(B_m\right)$ and average confidence $\mathrm{conf}\left(B_m\right)$ calculated bin-wise and weighted by the fraction of predictions present in each bin results in the ECE [16]. A commonly chosen value for M is 10 [17].

4.5.2 Uncertainty Calibration Error

The expected uncertainty calibration error (UCE) is interpreted as the weighted and binned average of the difference between the model error (the complement to the model accuracy) and the total uncertainty in the prediction (quantified commonly by the Shannon Entropy defined in Section 4.4). The UCE is given by

$$\begin{aligned}
\mathrm{UCE} &= \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{err}\left(B_m\right) - \mathrm{unc}\left(B_m\right)\right| \quad \text{with} \\
\mathrm{err}\left(B_m\right) &= 1/\left|B_m\right| \sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i \neq \mathbb{E}[\hat{y}_i]\right) \quad \text{and} \\
\mathrm{unc}\left(B_m\right) &= 1/\left|B_m\right| \sum_{i \in B_m} \mathcal{H}_i\left[\hat{\mathbf{p}}_i(\boldsymbol{y}_i|\boldsymbol{x}_i)\right] .
\end{aligned} \tag{4.14}$$

Here, $\mathrm{err}\left(B_m\right)$ denotes the error in each bin and $\mathrm{unc}\left(B_m\right)$ denotes the predictive uncertainty (also referred to as total uncertainty) in each bin.

4.5.3 Uncertainty Calibration Score

The Uncertainty Calibration Score is a relatively straightforward measure that estimates the gap in calibration. It is essentially derived from the area enclosed between the calibration curve of the model's predictions and the calibration curve for the perfectly calibrated case (i.e. the $y = x$ curve.) It was introduced by Wursthorn et al. for estimating uncertainty calibration in a regression setting (pose-estimation) [46]. The Uncertainty Calibration Score (UCS) is defined as

$$\mathrm{UCS} = 1 - \frac{A}{A_{max}} \,, \tag{4.15}$$

where A denotes the area enclosed between the calibration curve and the expected diagonal in case of perfect calibration. The largest enclosed area corresponds to the poorest case of calibration possible which is given by A_{max} = 0.25 [46]. In this work, as seen previously we attempt to estimate the calibration error through two similar yet complementary approaches using the ECE and the UCE. To use the UCS appropriately in both these approaches, it is proposed to redefine the UCS as the CCQS (Confidence Calibration Quality Score) and the UCQS (Uncertainty Calibration Quality Score). The CCQS is calculated from the reliability diagram for the ECE and the UCQS is calculated from the reliability diagram for the UCE respectively.

4.5.4 Area Under Sparsification Error

To 'sparsify' means to remove elements from an analysis. Sparsification refers to the removal of predictions from the evaluation set based on the ordering provided by an uncertainty estimation measure. A sparsification curve is computed as a function of sparsification. If an ideal uncertainty measure exists, then such a measure would indicate the highest uncertainty values for the wrongly classified pixels alone and would target the erroneous pixels first. Removing such erroneous predictions from the analysis immediately results in a higher accuracy. The sum of false posities and false negatively constitute the total error in the prediction. As soon as sparsification leads to the removal of all erroneous predictions, the evaluation is left with true positives alone and this leads to a maximum possible value of mIoU (or conversely the minimum possible value for an error based metric like the Brier score).

The following steps form the core of a sparsification analysis, which is generally performed for each class present in the evaluation set independently [47].

1. Compute the pixel-level uncertainty values of all pixels z (in an image or a batch) using an uncertainty estimator (like the previously discussed measures from Section 4.4).

2. Sort the uncertainty values obtained using the measure in the descending order.

3. Remove a fraction of the most uncertain pixels from the analysis (hence the term *sparsification*).

4. Remove the same fraction from the count of erroneous predictions (sum of false positives and false negatives) and calculate the oracle accuracy in terms of IoU (or error in terms of mean squared error).

5. Quantify the average prediction quality of the remaining pixels in the analysis (with a scoring rule or a DNN performance metric).

Steps 3, 4 and 5 are iteratively performed till there are no pixels to be removed from the analysis. A sparsification curve could be obtained in two ways from step 4.

- A scoring rule (for example a proper scoring rule like the Brier Score [48]) could be used to quantify the error in the prediction. This leads to a sparsification curve of the error vs. sparsification fraction. This approach has been used in the work of Gustafsson et al. [49].

- A robust DNN performance metric like the mIoU could be used to quantify the performance of the remaining pixels' prediction. This is the complement to the previous approach and this results in a sparsification curve of IoU vs. sparsification fraction. This approach has been used for LiDAR data successfully by Dreissig et al. [50]. They interpret the complement to the softmax likelihood as the predictive uncertainty measure - the Variation Ratio (V) [40], described in Section 4.4.

The upper bound of all such sparsification curves from a model is the oracle curve. Essentially, if a perfect "oracle" uncertainty measure exists, then that would target only the wrong predictions (namely the false-positives and false-negatives) and would have no impact on the pixels that are correctly classified by the model. That is why the oracle curve (if the analysis is evaluated based on mIoU for instance) is simply obtained by limiting the sparsification to the wrongly classified pixels alone.

Sparsification error is modeled as the difference between the oracle curve and the sparsification curve. An integral (over the complete domain, from no sparsification to complete sparsification) of the difference between these curves is referred to as the Area Under the Sparsification Curve (AUSE). While it is common to use the Shannon entropy from Equation (4.10) as an uncertainty measure to sort and subsequently sparsify the predictions, in this work in a later study in Section 5.1, the AUSE methodology has been implemented with the IoU as the evaluating metric and the Variation Ratio V as the sorting measure. This methodology has been found in literature too, in the work of Dreissig et al. [50]. An argument that supports the use of AUSE calculated using the Variation Ratio is that both the uncertainty measure and the evaluation metric are dependent only on the softmax likelihood i.e. the top-1 class in the prediction vector. This makes it similar to compare it with how the ECE

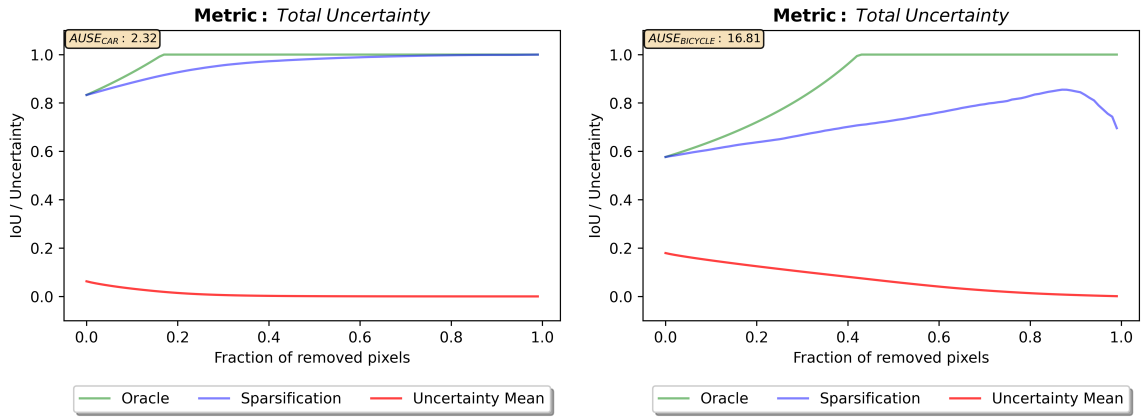

Figure 4.2: The oracle and the sparsification curve for two classes Car and $Bicycle$ are illustrated here. It is expected that the sparsification curve reaches the maximum (of 1.0, denoting 100% IoU) much later since there are a higher number of wrong predictions for a sparsely represented class like $Bicycle$ possibly because the model struggles to adequately learn the features of the class. For the $Bicycle$ class, the uncertainty measure indicates higher uncertainty values even for the correctly classified pixels towards the end of sparsification resulting in the drop in IoU. Here the total uncertainty refers to the Shannon entropy at a pixel-level.

is calculated in the reliability space.

In Figure (4.2), the oracle and the sparsification curves for two classes from the study on the KITTI dataset are illustrated. These two classes are largely different in terms of their representation in the chosen evaluation set of images.

5 Studies

Details about the studies carried out over the course of this work are provided in this section. The first study was performed to verify whether the hypotheses put forth by Wolf et al. [8] (also the broad objectives of this work from Section. 1.2) are valid when there is a shift in the input distribution of data due to aberrations from one dataset to another. The second study focused on the impact of temperature scaling as a post-hoc calibration approach for the outputs from a model. The third study focused on using the cross-entropy as the sorting measure and subsequently using the AUSE as an estimator of the residual uncertainty present in a model after training. The fourth and final study was carried out with Imatest Master, exploring the newly introduced Information Capacity metric. Additionally, the results obtained from Study 2 and 3 have also been submitted as part of a paper by me and Dominik Wolf as co-authors to the 2024 GCPR-VMV conference set to take place later this summer and the submitted manuscript is currently under review [16].

5.1 Study 1 - Inference on Cityscapes

In this study, an inference experiment using a pre-trained HRNetV2-W48 model was carried out on a perturbed set of images from two datasets of driving scenes from German streets, the KITTI and the Cityscapes [51, 52]. This was carried out to compare the inference of the network on perturbed versions of different datasets. The specifications and details of the experiment are presented subsequently. All the relevant data was recorded into a dataframe and later used for analysis. Every iteration, the chosen batch of images from both the datasets were convolved with the corresponding perturbed PSFs and then subsequently evaluated with the optical quality metrics (Section 4.1) and the neural network metrics (Sections 4.3 and 4.5) for each perturbed batch of images.

The HRNet (High Resolution Network) model architecture for dense perception tasks like semantic segmentation was introduced by Wang et al. [53]. The key feature of this architecture that makes it advantageous to use is that throughout the encoding process, features from multiple higher resolutions are maintained in parallel streams and subsequently fused. The main body of the network is composed of three components: multi-resolution convolutions, multi-resolution fusions and a representation head. A suitable resolution is determined for the high resolution convolution stream.

Typically the resolution of the input images are known and the images are passed into the main body through a stem which consists of two bi-strided convolutions with kernel dimensions 3×3 thereby reducing the resolution to $\frac{1}{4}$ of the original resolution. Downsampling of resolution is carried out in such a way that the resolution of a stream at an index r is reduced to $\frac{1}{2^{r-1}}$ of the resolution of the stream at the index

r = 1 (which is the first high resolution stream).

Wang et al. [53] introduce three variants with three different representation heads and refer to them as HRNetV1, HRNetV2, and HRNetV2p. HRNetV1 is trained for a pose estimation task on COCO, HRNetV2 is trained for a semantic segmentation task on the Cityscapes dataset and the HRNetV2p head is trained for an object/instance detection task on COCO [51], [52]. Specifically the HRNetV2 representation head has been used in this study for semantic segmentation with their respective details listed below [53]:

- V2 refers to the network head structure where the low resolution representations from the parallel streams are rescaled in size to match the resolution of the high resolution channel without changing the number of channels and concatenated together. This is followed by a 1 $\times$ 1 convolution to fuse the four representations (with resolutions $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$ respectively). The architecture is visualized in Figure 5.1 [53], [54].
- W48 refers to the width (number of channels) of the high resolution convolution at index r = 1.
- The trained model has 65.9 $\times 10^6$ parameters.
- The trained model performs at a computational cost of 696.2 GFLOPs (Giga Floating Point Operations Per Second) calculated at the full resolution of (1024 $\times$ 2048) [51].
- The model performs inference on the complete Cityscapes validation set at a single scale with an mIoU of 75.7%.

Although introduced by Microsoft Research, several variants of HRNet networks have been trained successfully by Google as part of their extensive study of training-transferability and have been subsequently uploaded to Kaggle [54]. It has to be emphasized while several models are available as pre-trained, the option to fine-tune them with further training is unavailable as the optimizer states (gradients) are not a part of their uploads.

The KITTI dataset developed by Geiger et al. [52] is one of the commonly used datasets for several dense perception tasks like object/instance detection and segmentation. KITTI provides semantic groundtruth labels for 200 images (from the streets of Karlsruhe, German city) for the training set and an additional 200 images for the test set for which the annotations are withheld for benchmarking purposes. Annotations are provided for 34 unique classes which are classified into 8 groups based on their semantic meaning/relevance. 19 critical classes that are also relatively well represented are used in the training of networks for semantic segmentation similar to the taxonomy devised by Cordts et al. [51] for Cityscapes.

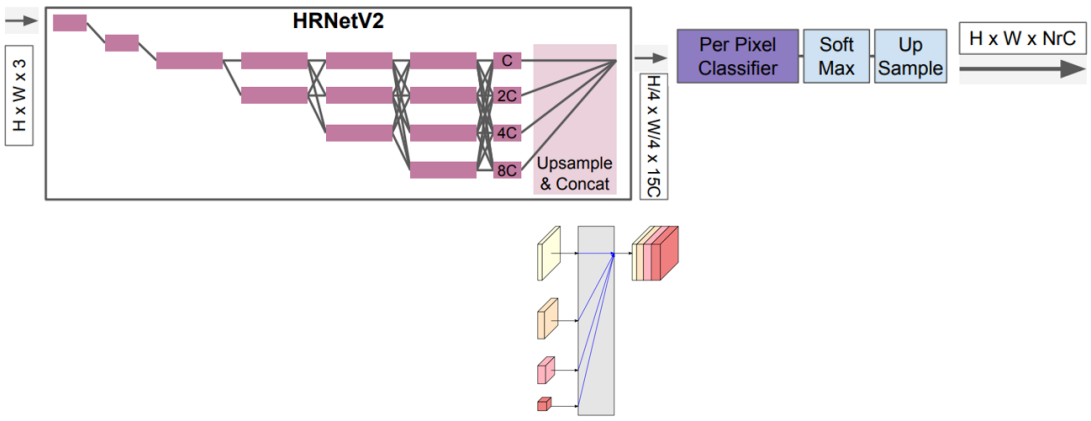

Figure 5.1: The representation head of HRNetV2-W48 used for semantic segmentation tasks. Illustrations combined from Wang et al. [53], Mensink et al. [54]

.

The Cityscapes dataset developed by Cordts et al. [51] is a very popular dataset for dense perception tasks specifically semantic and instance segmentation. There is greater diversity found in the driving scenes recorded in Cityscapes due to the data being recorded amidst varying degrees of road and pedestrian traffic. The Cityscapes dataset was recorded from a total of 50 German cities and offers 5000 densely annotated image-label pairs for training and validation. Typically, the validation split of 500 images set aside by Cordts et al.[51] is used for evaluation purposes. While the images are comparable to KITTI (since both datasets are from the streets of German cities), the resolution of Cityscapes images are substantially larger with a resolution of 1024×2048 pixels for all images.

5.2 Study 2 - Temperature scaling

The motivation to pursue a study on temperature scaling was that training neural networks which were considerably better in terms of calibration requires significantly higher computational resources. Due to computational constraints, the focus of this work turned towards exploring such post-hoc model independent calibration methods like temperature scaling. To begin with, temperature scaling's impact is studied on the evaluation set as a whole. This means considering all the classes present together and performing an all-at-once analysis. Then this study was extended by considering the impact of temperature scaling on individual classes. To study the impact of temperature scaling on different calibration measures, all the calibration measures discussed in Section 4.5 are employed namely the ECE, UCE, the UCS (with the proposed CCQS and the UCQS) and the counterparts from the sparsification space the AUSE_V and AUSE_S. This study used a customized UNET architecture [55] built by Wolf et al. [56] and trained on Audi's A2D2 driving dataset

as part of his doctoral studies. The model architecture comprises of five downsampling stages with the width of the network increasing by a factor of 2 before every downsampling step to retain information and reaches a maximum of 304 channels at the bottleneck of the UNET [55]. The depth of the network goes upto 48 layers excluding the input and output layer. Although the resolution of the A2D2 semantic segmentation dataset is 1208$\times$1920 pixels, keeping the computational resources in mind the resolution with which model training has been performed by Wolf et al.[56] is 604×960 pixels (a downsampling by a factor of 2). The trained UNET baseline that has been used in this study achieves a mIoU of 74.8% on the chosen evaluation set.

The evaluation set of data for this study consists of 200 images chosen randomly from the A2D2 semantic segmentation dataset. The A2D2 dataset provides annotations for 38 classes in its groundtruth data. But to remain consistent with the labeling taxonomy devised by Cordts et al.[51] for Cityscapes, the A2D2 groundtruth was relabeled aligning with the Cityscapes' annotations (which consists of 19 unique classes used for training and evaluation purposes).

5.3 Study 3 - AUSE in uncertainty estimation

For multiclass semantic segmentation, cross-entropy is a commonly used objective function. In a way cross-entropy attempts to quantify the similarity between a groundtruth label and its corresponding prediction from a model. If an optimal uncertainty measure is used in sorting the predictions of a model, the sparsification curve should align with the oracle curve, reducing the AUSE to zero. This alignment occurs because the oracle curve directly targets the sum of False Positives and False Negatives in the predictions. Therefore, we suggest using the AUSE, derived when cross-entropy is used as the uncertainty measure for sorting predictions, as an estimator of the residual uncertainty. This residual uncertainty is non-reducible when the network architecture is fixed and includes the aleatoric uncertainty from the data generation process as well as the model uncertainty due to the constraints of the chosen neural network architecture. This assumption holds if the model training converges. As training progresses and before overfitting occurs, the epistemic uncertainty—comprising both approximation uncertainty and model uncertainty [57]—is expected to decrease.

To test this hypothesis, we trained five models for an increasing number of epochs, keeping the model architecture and hyperparameters constant. The five models corresponded to 50, 100, 200, 300 and 400 epochs respectively. To be considerate of the role of epistemic uncertainty (which is also due to the uncertainty in the model's weights, the 400-epoch model used in the previous study is not used but a new model is trained. Nevertheless, it was cautiously observed if overfitting was being reached as a byproduct of extended training. Simultaneously the frequencies of

occurrence of the various classes in the evaluation set was also recorded. The most frequently occurring classes in the selected evaluation set were used to assess the AUSE with cross-entropy ($AUSE_{CE}$).

5.4 Study 4 - Information Capacity

A preliminary investigation of the nature of Information Capacity, as an image quality metric was carried out over the course of this work. To this end, a set of five slanted-edge images were simulated artificially by one of Volkswagen's Tier-1 suppliers. A temporary license for Imatest LLC's proprietary image quality software Imatest Master was obtained officially for this study. These images were simulated with varying degrees of noise and produced different MTFs when passed through the in-house e-SFR algorithm. It has to be emphasized that the nature and details of the simulated noise is unknown and thus the noise was assumed as black-box noise. The following steps formed a major part of this study to measure the information capacity of these five images.

1. The spatial frequency responses produced by Imatest Master were first validated using two comparable e-SFR implementations - one from Volkswagen AG and the other from the Tier-1 supplier that supplied the images in the first place.

2. To ensure comparability of results, the pre-processing steps taken were kept consistent, for example the histogram optimization to ensure balance in the dark:bright ratio and the demosaicing method applied onto the raw image.

3. Once the MTF results were proven to be consistent, the other sequential results were calculated using Imatest Master - for example the noise image followed by the NPS, NEQ and finally the Information Capacity value C.

4. It needs to be emphasized that there is no alternative to Imatest's implementation of the Information Capacity algorithm described in detail in Algorithm 1. Due to this, only the general behaviour of Information Capacity as an optical quality indicator could be studied.

6 Results and Discussion

This chapter lists down the key results obtained from the studies carried out through the course of this work. The results for the four studies are presented here in the same order as that of Section 5.

6.1 Results - Study 1

The initial objective was to prove if the robustness of the HRNetV2 model to perturbations added on images from the KITTI dataset was reproducible and independent of the image-batch chosen. Wolf et al. [8] used 40 images as a batch from the KITTI semantic segmentation dataset to carry out studies with their optical threat models. In my study, from the training set provided, an evaluation dataset of 50 images were randomly sampled for the study and perturbations were added to the batch at a Zernike-coefficient cardinality of 125. This cardinality corresponds to the three coefficients ω_3, ω_4 and ω_5 being equidistantly sampled at -1.0, -0.5, 0, +0.5 and +1.0. To avoid discrepancies in the image resolution, all images were cropped to match the smallest image in the KITTI training dataset for semantic segmentation (370 $\times$ 1224). Figure 6.1 illustrates the impact of perturbations from the optical threat model on an image from the KITTI dataset.

The occurrence of classes in the chosen evaluation set are illustrated in Figure 6.2. Before a choice on the size of the evaluation set was made, model inference was performed sequentially on increasing batch sizes to observe if there was a significant impact on the batch size chosen. In Figure 6.3, this is illustrated. Keeping the already chosen batch size of 40 by Wolf et al. [8] and the computational load of having a larger batch size in mind, the batch size was chosen to be 50.

Since the resolution of the Cityscapes images (1024 $\times$ 2048 pixels) are significantly larger than that of KITTI, the batch size had to be appropriately reduced (to 30 images) considering the computational overhead. The HRNetV2 model takes in a batch of images as a tensor of shape N$\times$H$\times$W$\times$3 and produces a prediction tensor of the shape N$\times$H$\times$W$\times$34 indicating the 34 classes present in the complete dataset, where N, H and W correspond to the batch size, height and the width of the images. Figure 6.4 illustrates the occurrence of classes present in the chosen evaluation set.

6.1.1 Bijectivity

The functional relationships between metrics that characterize optical quality and metrics that characterize DNN performance were compared between the study on KITTI and Cityscapes respectively. It was observed that the relationship with the neural network metrics (mIoU, mECE) for refractive power and MTF calculated at the half-Nyquist frequency were similar and consistent for KITTI and Cityscapes. In

Figure 6.1: The row on the top is raw image before and after the addition of the perturbations parameterized by the optical threat model. A zoomed-in 200 $\times$ 200 pixels region of interest containing a car makes it evident the adverse impact of aberrations on the information contents of an object in an image.

both the studies, a similar pattern emerged and it was not possible to unambiguously infer an mIoU and mECE value from each unique refractive power and MTF value pointing to pitfalls in using refractive power and MTF as a optical quality indicator. Figure 6.5 illustrates this observation.

A similar comparison for the Strehl ratio (SR) and the Optical informative gain (OIG) is illustrated in Figure 6.6.

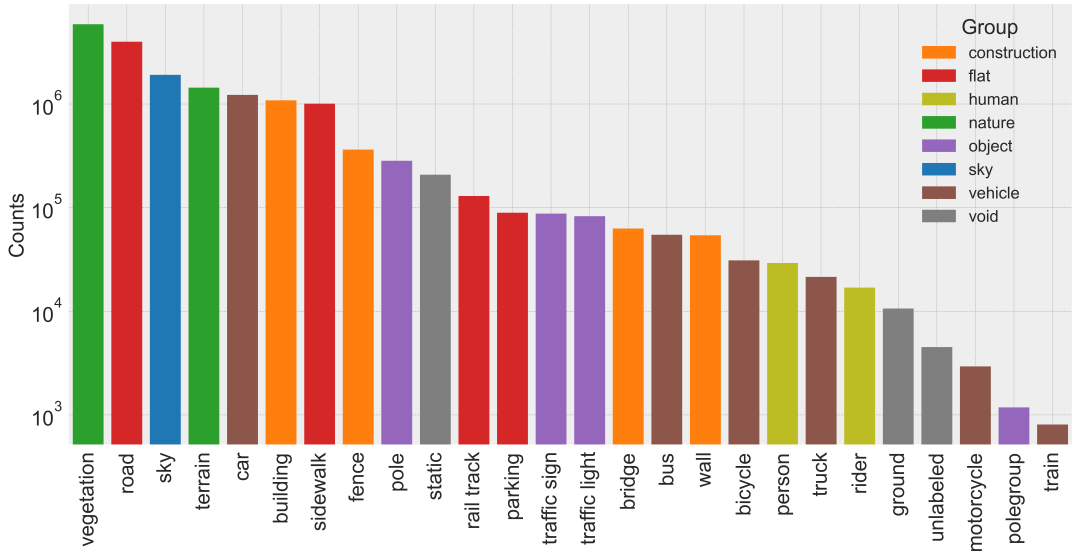

Figure 6.2: The occurrence of classes (each belonging to one of the 8 groups) is sorted from the most represented to the least represented class in the chosen evaluation set from KITTI is illustrated here.

6.1.2 Relative impact of pure defocus

From the work of Wolf et al. [8], the expectation was that the pure defocus optical effect (arising from the effects of Zernike coefficient ω_4) dominated over the effects of ω_3 and ω_5 causing astigmatic aberrations. This was also similarly observed in both the analyses and presented in Figure 6.7. The sensitivities of the metrics that have been employed to the three Zernike coefficient features are quantified using the highly intuitive and effective Shapley values [58]. Shapley is a local model-agnostic explanation method that has found a lot of relevance in explainable AI [59]. Briefly described in Equation (6.1), the Shapley value ϕ for a feature i and an objective function $\hat{\Xi}$ is determined for a particular feature subset $\mathcal{M}_f$ as

$$\varphi_i(\hat{\Xi}) := \sum_{\mathcal{S} \subseteq \mathcal{M}_f \setminus \{i\}} \binom{|\mathcal{M}_f| - 1}{|\mathcal{S}|}^{-1} \frac{[\hat{\Xi}(\mathcal{S} \cup \{i\}) - \hat{\Xi}(\mathcal{S})]}{|\mathcal{M}_f|}. \tag{6.1}$$

The Shapley values are determined by weighting the individual coalition quantity with $[\binom{|\mathcal{M}_f|-1}{|\mathcal{S}|}]^{-1}$, the inverse of the binomial coefficient. A detailed explanation and its theoretical founding is provided in the work of Shapley et al. [58]. In this study, for easier visualization, the Shapley values are plotted in a normalized fashion (normalized to the effect of ω_4 but the unnormalized Shapley contribution values are recorded too. In Figure A.1 in Section A.3, the Shapley plots for the four optical quality metrics are illustrated, which further indicate the dominant impact of ω_4.

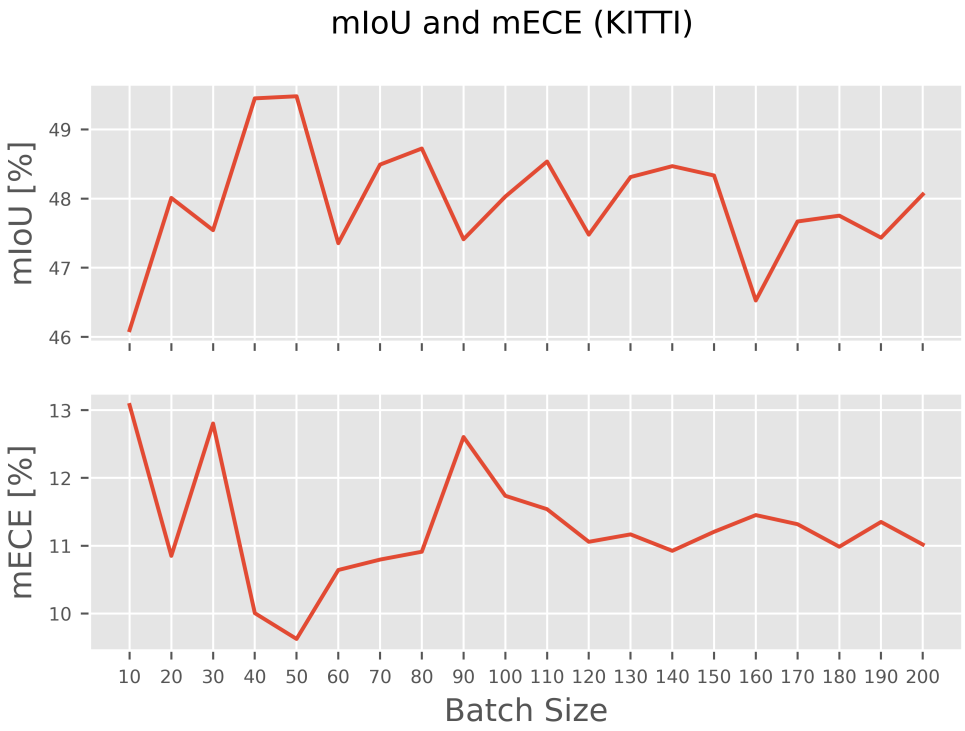

Figure 6.3: The ΔmIoU and ΔmECE between analyses with batchsize 10 and 20 is +1.9% and -2.2% respectively. Nevertheless, the computational overhead is much higher when a larger batchsize is chosen (for example $>$ 100). Hence 50 randomly sampled images are chosen for the analysis.

6.1.3 Calibration

The mECE is found in both the analyses and presented in the reliability diagrams illustrated in Figure 6.8 along with their respective baseline mECE values (for the diffraction-limited case). It has to be emphasised that from the available literature behind the model's training by Google [54], no conclusive evidence could be found about a training-time augmentation method or a post-hoc method used to calibrate the prediction vectors of the model. Hence, a calibration temperature T = 1.0 is assumed that for both the reliability diagrams.

From Figure 6.8, it is clearly evident that for both the datasets the model was overconfident (the bins where the histogram falls short of the perfect calibration curve $y = x$). This provided sufficient motivation to pursue studies in the direction of post-hoc calibration of model outputs. An additional analysis was performed with the AUSE methodology. The equivalent neighbor of ECE in the sparsification space AUSE_V (AUSE calculated using the Variation ratio V as the sorting measure) was chosen for the analysis and was calculated at every iteration (every Zernike coefficients set.) The sensitivity of AUSE to the Zernike coefficients was also similar to ECE such that the pure defocus was found to be dominating. However, this analy-

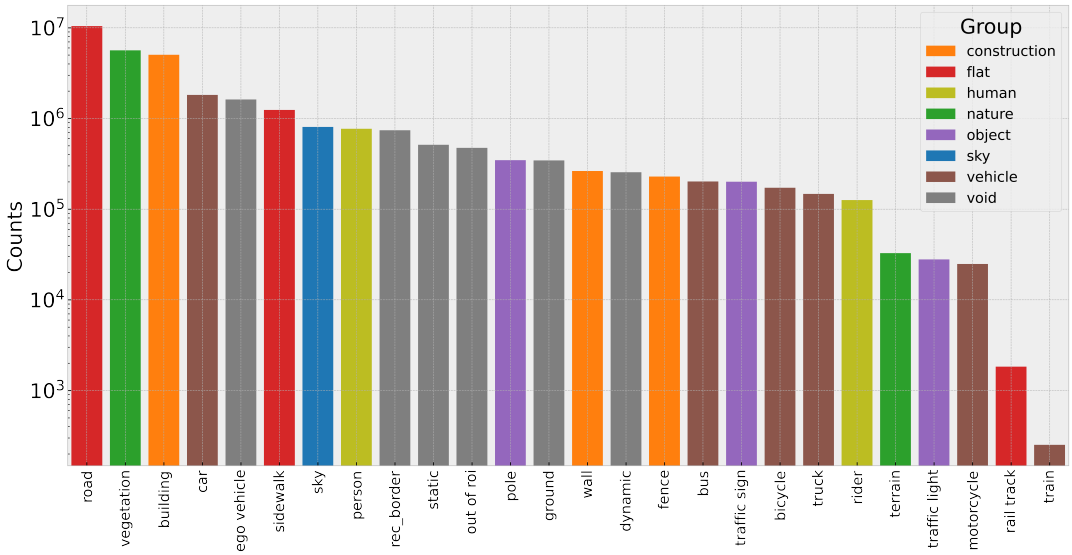

Figure 6.4: The occurrence of classes (each belonging to one of the 8 groups) is sorted from the most represented to the least represented class in the chosen evaluation set from Cityscapes is illustrated here.

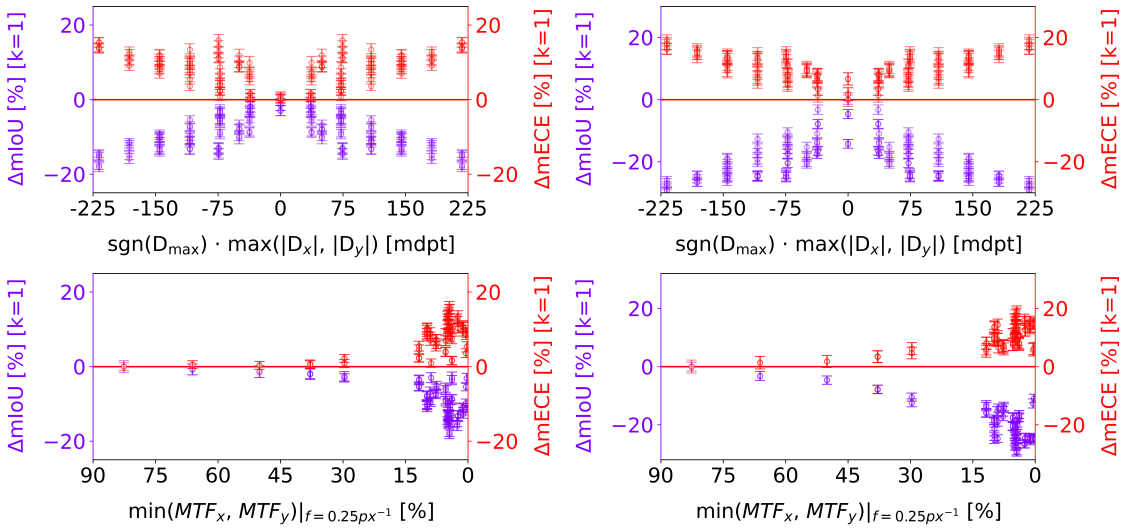

Figure 6.5: (L): The column on the left corresponds to the refractive power and MTF calculated at the half-Nyquist frequency for the KITTI based study. The peak adverse impact on the mIoU caused by the maximum perturbation case ($|\omega_3|$=$|\omega_4|$=$|\omega_5|$=1) is found to be ΔmIoU = -19.3% and for the mECE it is ΔmECE is +20.2% respectively. (R): The column on the right corresponds to the refractive power and MTF calculated at the half-Nyquist frequency for the Cityscapes study. A similar behaviour is observed, with the impact on the mIoU noticeably higher in this study. The peak adverse impact caused by the maximum perturbation case ($|\omega_3|$=$|\omega_4|$=$|\omega_5|$=1) for the mIoU is ΔmIoU = -32.8% and for the mECE it is ΔmECE = +19.8% respectively.

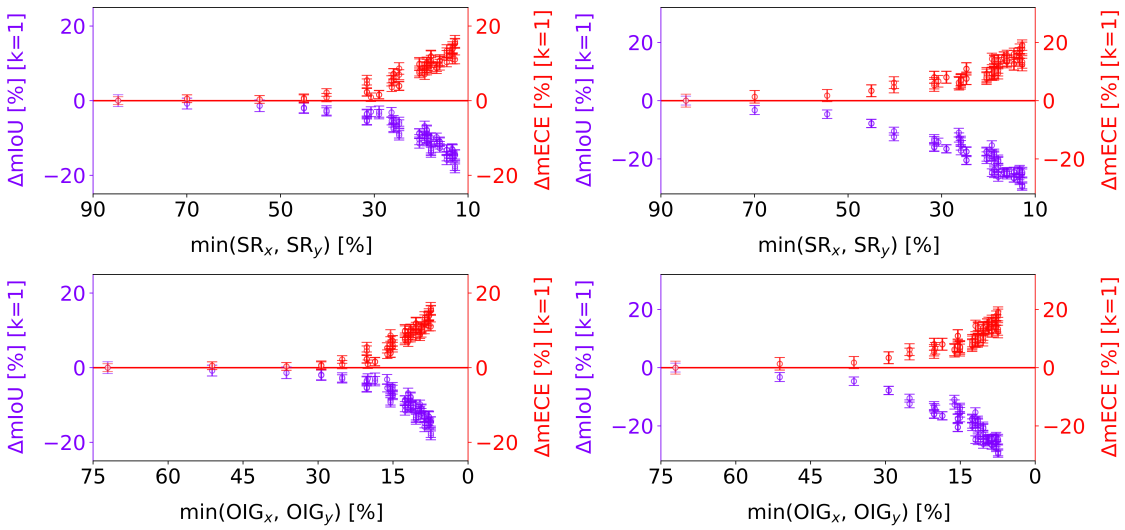

Figure 6.6: (L): The column on the left corresponds to the Strehl Ratio and OIG calculated for the KITTI based study. (R): The column on the right corresponds to the Strehl ratio and OIG for the Cityscapes study. There exists a functional relationship for the Strehl ratio and OIG with the DNN metrics within the uncertainty intervals for mIoU and mECE. The uncertainty bars presented in the plots here and from Figure 6.5 are the standard deviation of the mean of mIoU and mECE calculated for all the images present in the analysis (50 for KITTI and 30 for Cityscapes respectively).

sis was performed only on classes that are well represented. High representation equates to more occurrences in the analysis (Figure 6.2 and Figure 6.4). Additionally the Area Under Mean Uncertainty (AUMU) is also calculated. Mean uncertainty in this case refers to the average over all pixel-level Variation ratio values in an image. The results for the class Car from the Cityscapes analysis are illustrated in Figure 6.9.

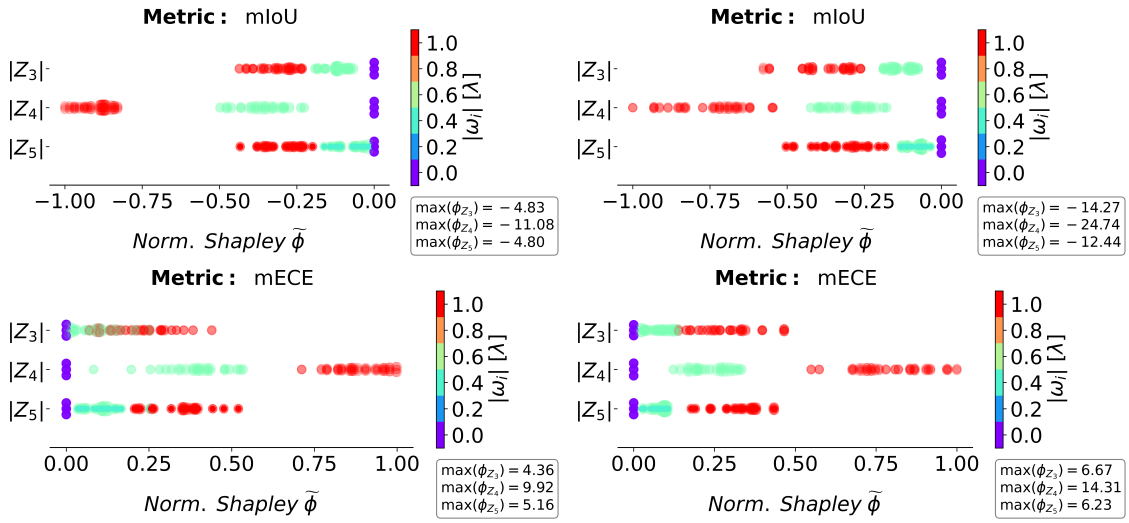

Figure 6.7: (L): The column on the left corresponds to the sensitivity of mIoU and mECE to the three Zernike coefficients for the KITTI based study. (R): The column on the right corresponds to the sensitivity of mIoU and mECE to the three Zernike coefficients for the Cityscapes based study. The sensitivities are quantified with the help of Shapley values, which are normalized by the effect of ω_4 and presented in this figure. It is evident that pure defocus ω_4 dominates over astigmatic aberrations and the individual unnormalized Shapley feature contribution values are different between the KITTI and Cityscapes studies.

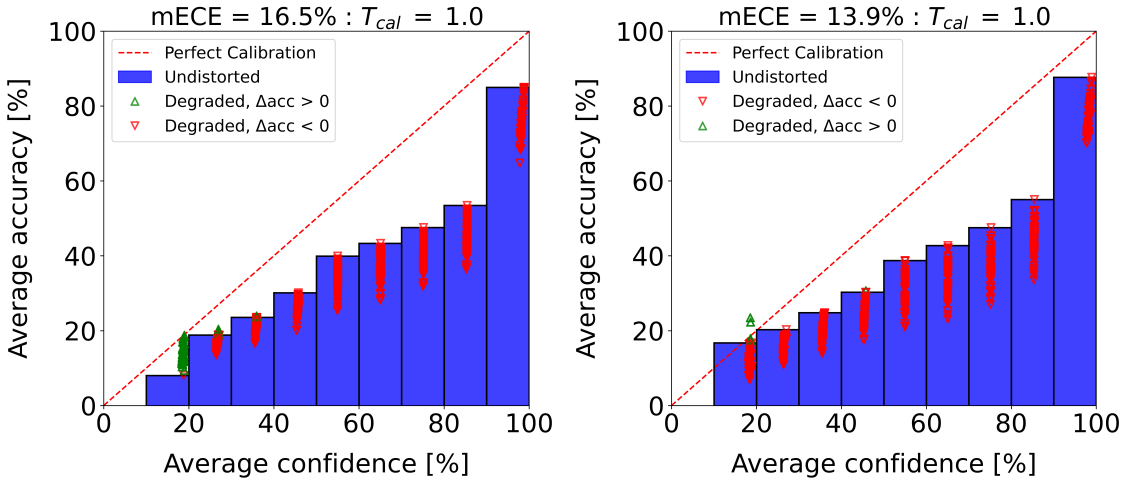

Figure 6.8: The reliability diagram on the left corresponds to the KITTI study and the one on the right corresponds to the Cityscapes study. The green bars present in both diagrams' second confidence bins $\in (10\%, 20\%]$ indicate a counter-intuitive artefact: it indicates that the degradation of the images resulted in a minor improvement in accuracy for the pixels binned into that interval. But it has to be understood as an artefact considering the already poor confidence of the model generally found at class boundaries [8]. The red bars answer to the expectation that aberrations degrade the model's calibration.

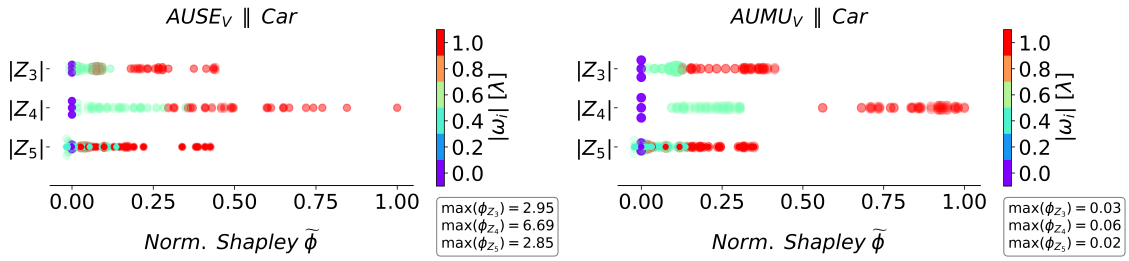

Figure 6.9: The pure defocus term clearly dominates in terms of impact on both the AUSE$_V$ an AUMU$_V$. It was observed in the study that since each analysis comprises of millions of pixels taken together as a batch, the AUMU$_V$ had a high inertia i.e. even a huge drop in the AUSE value only brought about a very minor degradation of the AUMU value.

6.2 Results - Study 2

All calibration measures discussed in Section 4.5, if they were estimating the true miscalibration of the probabilities, should have been minimized by the same temperature parameter. We observe this is not the case. Figure 6.10 shows the impact of temperature on the metrics based on reliability-diagrams namely (ECE, CCQS, UCE and UCQS). The optimal temperature was observed to be 0.4 for the complete evaluation set of 200 images. The ECE and UCE were both minimized at this temperature. The distribution of data is concentrated in just one bin - the first bin corresponding to $[0\%, 10\%]$ uncertainty for UCE and the last bin corresponding to $(90\%, 100\%]$ confidence for the ECE. Since the ECE/UCE is a weighted average, minimizing the gap in calibration in this bin alone sufficiently leads to the minimization of ECE/UCE over the entire range. But this is not the case for minimizing the enclosed areas for CCQS and UCQS. Here, for the areas to be minimum, all the bins' miscalibration have to be simultaneously minimized. This is why it can be noticed that the enclosed areas for the optimal temperature $T_{\text{cal}} = 0.4$ for ECE/UCE are not the minimum areas. The optimal temperature if the areas A_{CC} and A_{UC} are to be minimized are 0.4 and 0.9 respectively.

It is noticeable that the predictions in the first bin (for the UCE) and the last bin (for the ECE) dominate the ECE/UCE values and thereby the temperature scaling process too, exposing a problem with binning based calibration measures, especially when the data is extremely concentrated at one or two bins. The CCQS and UCQS strive to achieve the case of perfect calibration not just weighted by one bin but throughout the range of confidences and accuracies and hence do not behave similar to ECE/UCE.

Figure 6.11 further illustrates the impact of the optimal temperature on miscalibration (gap between average accuracy and confidence) in each bin along with histograms indicating the number of predictions in each bin.

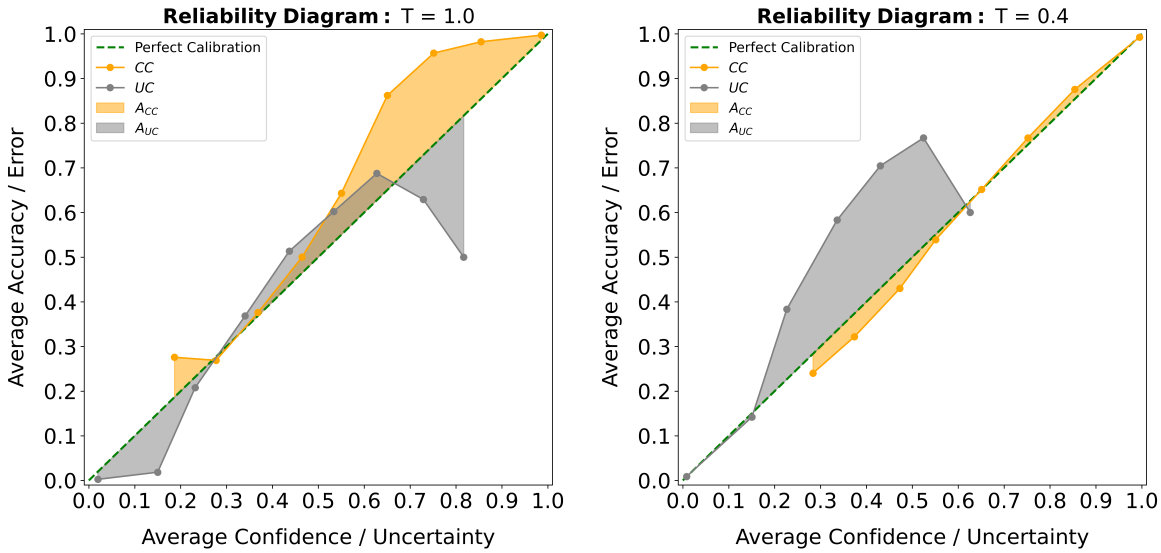

Figure 6.10: The reliability diagrams for the untempered case (T = 1.0, left) and for the optimal temperature (right) are illustrated along with the enclosed areas A_{CC} and A_{UC}.

An analysis of how temperature scaling impacts each class present in the evaluation set is interesting and essential because the learning of a class's features depend on the level of representation available in the inputs and thus calibration is also dependent on the particular class's representation. The metrics considered for a class-wise analysis are: the ECE (conventionally considering only the top-1 softmax confidence), the UCE (using Shannon entropy) and its two equivalent metrics from the sparsification space, the AUSE calculated using variation ratio as the sorting measure (AUSE$_V$) and the AUSE calculated using Shannon entropy as the sorting measure (AUSE$_S$) respectively. The IoU was utilized to evaluate the sparsified evaluation set pertaining to each class. The optimal temperature that minimizes these four metrics differs from one class to another. This discrepancy is due to the relative differences in the level of representation of each class in the data available to the model during training. In Table 6.1, the optimal temperatures T_{cal} for the most represented and the least represented classes (three classes from both groups) in the evaluation set are tabulated.

6.2.1 Average impact of temperature scaling

As an extension from the class-specific impact of temperature scaling on calibration to understanding the overall impact on all classes in the evaluation set, the mean of each metric for each class is calculated across an equidistant range of temperatures $T \in [0.1, 10]$, including the NLL and the Brier score [25]. For example, the class-specific ECEs at each temperature was calculated and their averages were plotted. It needs to be highlighted that there is a distinction between calculating the ECE by considering all classes together at once and averaging the class-specific ECE val-

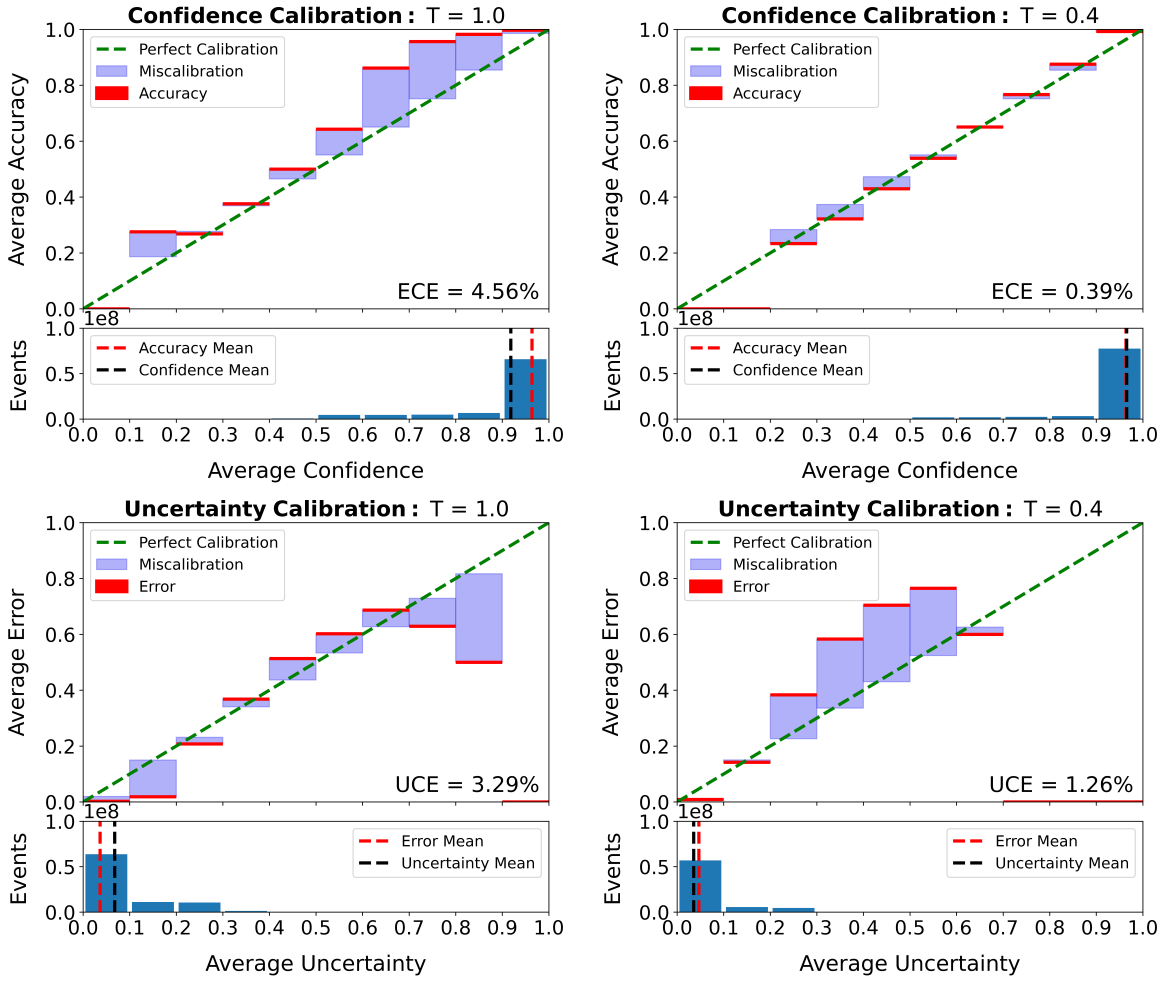

Figure 6.11: Calibration does not guarantee that the miscalibration in every bin is minimized since the bin-wise distribution of predictions is extremely skewed. This provides evidence on why the CCQS and UCQS for skewed data will not be coherently maximized as the ECE and the UCE are minimized.

Class ID	Class	Reliability		Sparsification	
		$\underset{T}{\text{argmin}}\ \text{ECE}$	$\underset{T}{\text{argmin}}\ \text{UCE}$	$\underset{T}{\text{argmin}}\ \text{AUSE}_V$	$\underset{T}{\text{argmin}}\ \text{AUSE}_S$
7	Nature	0.4	0.4	1.8	0.9
8	Sky	0.4	0.4	1.4	0.9
0	Road	0.5	0.5	1.7	0.1
9	Pedestrian	0.3	0.4	0.9	0.9
12	Small Vehicle	0.2	0.3	1.9	1.0
15	Animal	0.9	1.2	0.7	0.6

Table 6.1: T_{cal}, which minimizes the respective calibration metric, is listed for the three most represented (top three classes) and the three least represented classes (bottom three classes) in the evaluation set. The standard deviation of the class-wise calibration temperature is higher in the case of less represented classes like $Animal$.

ues, as shown by Kull et al.[60].

Since the AUSE methodology in our study employs the IoU for evaluation, a class-specific evaluation is necessary to address class imbalances, which could lead to an anomaly when the oracle overshoots the sparsification curve. The reason behind this is that when the wrongly classified pixels of a particular class are sequentially removed, the IoU is monotonically increasing. But for the mIoU when all classes are considered this is not the case since a constant sparsification amount has different impacts on different classes. This is illustrated in a simple experiment in Table A.1 in Appendix A. An average of class-wise sparsification studies eliminates this issue. Additionally, this study reveals that the optimal temperature for the ECE/UCE, when calculated including all classes simultaneously, is the same as that for the mean ECE/UCE obtained by averaging over all classes. The average performances of the calibration measures across temperatures are summarized in Figure6.12.

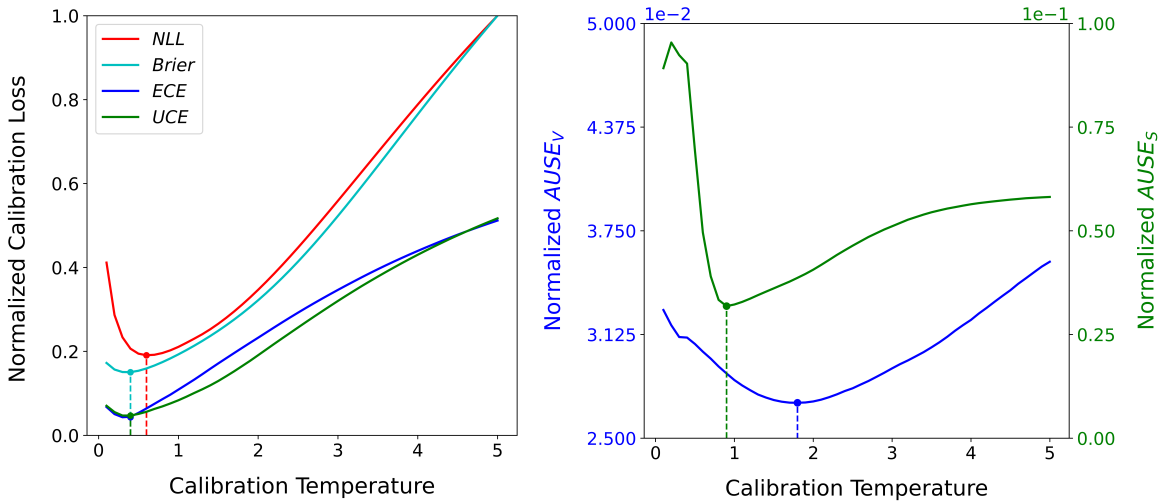

Figure 6.12: Left: The normalized calibration loss surface of NLL, Brier score, the mean of class-wise ECE and the UCE are plotted. The optimal temperature for the Brier score, ECE and UCE coincide at T = 0.4 while the NLL indicates a minimum at T = 0.6. Right: The normalized calibration loss surface of AUSE_V and AUSE_S indicate minima at T=1.8 and T=0.9 respectively.

6.3 Results - Study 3

The performance of the five models used in this study were increasing from the least trained model to the most trained model. On the same 200 image A2D2 evaluation set, the five models resulted in a 64.5%, 67.6%, 70.8%, 72.3% and 74.8% mIoU. This helped in ruling out the possibility of overfitting. The reason why the inadequately

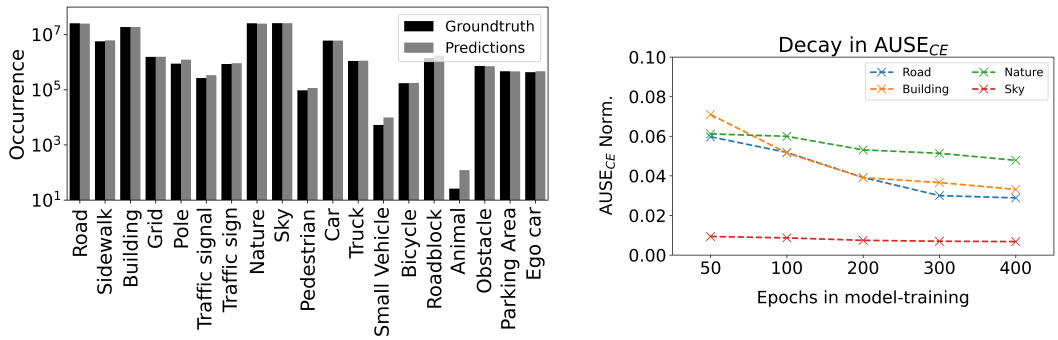

Figure 6.13: From the evaluation set considered, the four classes with an occurrence of $> 10^7$ are chosen for the analysis. It is noticeable that the AUSE$_{CE}$ decays to the smallest value for the best-trained 400-epoch model for each of these classes.

represented classes are avoided in this study is because for the lesser trained models, there is no guarantee on whether the class's features were learned or not especially for the insufficiently trained models. Hence the mIoU does not really reflect the level of learning for such classes. In Figure 6.13, the decay of AUSE$_{CE}$ for the top four classes is illustrated. This provides an interesting result that the AUSE methodology could be used in estimating the residual uncertainty in a model.

6.4 Results - Study 4

As discussed in Section 5.4, the results obtained by the three comparable ISO:12233 based spatial frequency responses are first discussed. The five images received from the Tier-1 supplier were all of the size 100 $\times$ 100 pixels and contained exactly one slanted edge with an angle $\alpha \in [5°, 10°]$. In Figure 6.14, the resulting MTFs from the different implementations are illustrated.

In Figure 6.15, the plots of the Noise Power Spectrum NPS and the Noise Equivalent Quanta NEQ are illustrated. It can be clearly noticed that the power spectrum of noise in each image is different.

A conclusion that can be obtained from the Information Capacity results in Table 6.2 and the study as such is that the Information Capacity as a metric correlates in a similar manner to the spatial frequency response i.e when the image is deteriorated by noise contents, the Information Capacity is also affected negatively similar to the MTF.

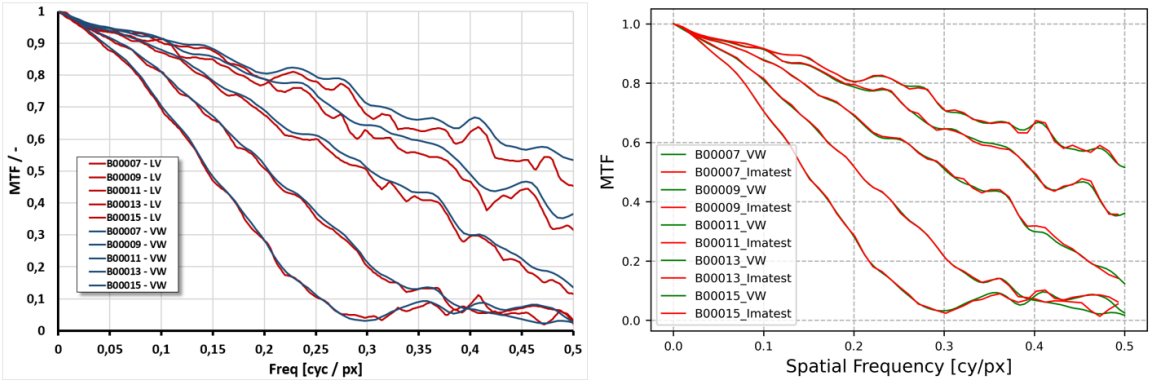

Figure 6.14: The figure on the left, provided by the Tier-1 supplier compares the difference between their e-SFR implementation and the in-house implementation in terms of the MTF (normalized to 1). Hence the figure is also visualized differently. The figure on the right compares the results from Imatest Master with the in-house results. The difference between ours and Imatest's is significantly smaller throughout the domain (0 to Nyquist frequency)

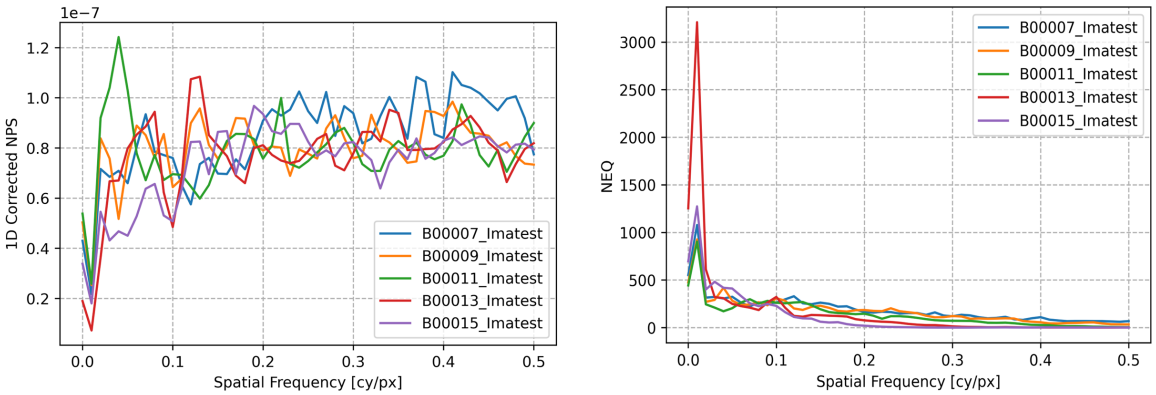

Figure 6.15: The NPS and the NEQ are dimensionless quantities indicating a number. As illustrated on the left the radially averaged and normalized NPS is different at different spatial frequencies for each image. While this indicates that the noise components are varying across frequencies, it is not sufficient to deduce what kind of noise these spectra relate to.

Image ID	MTF ($\frac{Ny}{2}$) [%]	C_{NEQ} [b/px]	C_{max} [b/px]
B00007	79.961	2.819	7.286
B00009	74.192	2.699	7.145
B00011	61.367	2.363	6.741
B00013	36.912	1.759	5.803
B00015	11.246	1.368	5.261

Table 6.2: C_{NEQ} and C_{max} results for the five simulated images, sorted from the image with the lowest noise to the highest noise.

7 Limitations

This is an attempt to briefly outline the limitations I faced through the course of this work.

- The lack of driving-scene datasets that did not include a windshield in front of the camera is one limitation, not just for this work but in general for a study involving an additional lens-like element on the optical path (i.e. the windshield). Any image contains noise sourced from different components of the imaging system like the camera itself, the imaging sensor etc. For the simulated optical aberrations to be added to the images of an open-source dataset, the optimal case would be when the dataset originally was captured not from within a car but with the cameras being entirely exposed. But this is true, to the best of my knowledge only for KITTI and A2D2 [52, 61]. The study on Cityscapes was nevertheless performed with the assumption that the impact of optical aberrations from the original windshield on the Cityscapes car is negligible because the camera used for the Cityscapes dataset has a sufficiently low pixel resolution per field angle.)

- Computational overhead: One of the larger challenges faced during the course of this study was to manage GPU resources. Computer vision is resource-intensive as such with image-based studies being difficult and infeasible to scale up. Initially this study was planned to include inference studies on state-of-the-art vision transformers (ViTs) to study the impact of perturbations on vision transformers that have a different architecture as compared to that of a convolutional network. But this could not be completed due to the large computational requirements (GPUs) to train SOTA transformer architectures and work with them.

- Availability of pre-trained models: Although the computer vision community online is better than most fields of time, it is still limiting in the sense that most models which have been trained by a research group do not have the gradients (optimizer states) uploaded online along with the weights. This facilitates only an inference study but not fine-tuning of a network with further (augmented) data. I found this limitation in many models uploaded online in the respective Github repositories.

- The optical threat model developed by Wolf et al. [8] in his doctoral work limited the consideration to second-order Zernike coefficients. While this was a conscious and physically motivated decision, a threat model incorporating higher order effects like coma, trefoil and tetrafoil too could lead to either new and interesting results or prove the hypothesis that the much higher order terms need not be considered.

8 Future Work

I find multiple directions in which further research could be pursued as a continuation of these studies. This chapter briefly looks into the open questions which could be answered by future work both in the short-term and long-term.

8.1 Short term

1. The optical threat model developed by Wolf et al. [8] is specifically tuned for long focal-length telephoto lenses. This optical threat model need to be adapted by accomodating the field angle into consideration. Me and our group finds this to be the next possible valuable extension to the current version of the model. This extension would mean that the optical threat model could be tuned for a camera with a wide FoV.

2. Inference studies that have been performed to study the robustness of a model to optical aberrations could be extended to state of the art vision transformer architectures since recent research indicates that transformers are comparatively robust to local perturbations on an image [62].

3. With adequate support from the development team of Imatest LLC, the in-house codebase to calculate Information Capacity using Algorithm 1 can be validated comparing the results obtained from Imatest Master. This could lead to an in-depth understanding of the metric's behaviour.

8.2 Long term

1. The ultimate goal of this sub-domain of autonomous vehicle development is that the impact of optical aberrations induced by the windshield on the entire sensor-fusion chain should be studied. This would mean that a winshield has a quality threshold that encompasses its impact on data from the complete sensor-suite (aberrations on camera-based images, distortions on LiDar point clouds etc.) This, in my opinion could help one learn what is the total impact of a windshield on the entire perception process.

2. Another interesting approach to extend the study in the direction of understanding miscalibration is to develop bounds for miscalibration when windshields induce optical aberrations on the images. This is currently being worked on for adversarial attacks and referred to as certified robustness of a model to an adversarial attack [63]. In such a study, a model is essentially proven to be robust against an adversarial attack up to a certain norm (for example, an l_2 ball). A similar study on certifying calibration error and finding the theoretical bounds for calibration for a model has already been published by Emde et al. [64].

9 Conclusion

The concluding remarks have been addressed separately to each of the four major studies from this work from Section 5.

1. The impact of optical aberrations induced by the two parameterizations of the optical threat model - one for the KITTI and another for the Cityscapes dataset are similar and comparable. Additionally, while the individual impacts between the two datasets are different, the defocus is the dominating optical effect in both cases too. In fact, through the course of these studies, Wolf et al. [65] also observed during his work a comparable impact on the robustness on a UNET model trained on A2D2 giving sufficient proof that the impact on different datasets and architectures are comparable. The question of whether similar optical threat models from other sources have the same impact on the calibration and robustness even on completely different model architectures is still open.

2. With the study on the impact of temperature scaling, I believe it has been sufficiently proven that there exists a decoupling between different calibration approaches. With a focus on class-specific analyses, it is also observed that the optimal temperature for different classes lie at different locations. This leads us to understand that a two-fold requirement from the state of ideal calibration and ideal temperature is present: each class should be calibrated and each bin (from the reliability diagram) should be calibrated simultaneously approaching the theoretical expectation of zero calibration error.

3. The study of AUSE with cross-entropy used as the uncertainty measure led to an interesting observation that the AUSE (in this case acting as an estimator of the residual uncertainty) decays with increasing model training. Extended work on this could lead to yet another piece in solving the puzzle of quantifying the multiple components of predictive uncertainty.

4. While an attempt has been made to study the behaviour of Information Capacity, nothing conclusive can be declared yet. A general correlation between MTF and Information Capacity to the extent that noisy images that produce a poor MTF also produce a reduced Information Capacity value. But the exact functional relationship between noise, spatial resolution and this metric remains to be analysed.

My concluding remarks as a result of this work is that the impact of optical aberrations induced by the windshield on the robustness of AI models is significant, especially for a highly safety-critical application like autonomous driving. Any adverse impact on the calibration of a model's outputs demands to be considered seriously as they pose a direct threat to the safe operation of an autonomous system. While the impact on model accuracy and calibration has been quantified and compared through studies in this work, the open question remains: which is the best possible metric that sufficiently characterizes optical quality and also displays a bijective relationship with the DNN performance metrics (like mIoU, mECE etc.)? Further work on the behaviour of Information Capacity as a potential optical quality indicator could result in conclusive evidence whether it is in any way superior to the metrics already formulated and studied in detail by Wolf et al. [8].

While studying the impact of temperature scaling or the decoupling between different calibration measures was not a part of the initial goals of this project, it helped in understanding how there are a lot of questions yet to be answered about why miscalibration occurs in the first place and how it can be accurately estimated with the multitude of measures available. Only this will lead to a holistic solution to completely alleviate miscalibration (both under-confidence and over-confidence) in AI models especially working on safety critical domains like autonomous driving.

A Appendix

A.1 Jensen's Inequality

Equation (4.10) is simplified and rewritten here in terms of z as

$$\mathcal{H}\left[\hat{\mathbf{p}}(z)\right] = -\frac{1}{\log\left(K\right)} \sum_{i=1}^{K} \hat{p}_i(z) \cdot \log\left(\hat{p}_i(z)\right) \ . \tag{A.1}$$

Jensen's inequality states that for an arbitrary convex function ξ,

$$\xi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \leq \frac{\sum a_i \xi\left(x_i\right)}{\sum a_i}. \tag{A.2}$$

The inequality is reversed in case of a concave function ς,

$$\varsigma\left(\frac{\sum a_i x_i}{\sum a_i}\right) \geq \frac{\sum a_i \varsigma\left(x_i\right)}{\sum a_i}. \tag{A.3}$$

Substituting a_i with $p(z)$ and ς with $\log(z)$, the required result is obtained for the case of all elements of the probability vector being equal:

$$\begin{aligned} \mathcal{H}\left[\hat{\mathbf{p}}(z)\right] &= -\sum_{i=1}^{K} \hat{p}(z) \cdot \log(\hat{p}(z)) \\ &= \sum_{i=1}^{K} \hat{p}(z) \cdot \log(\frac{1}{\hat{p}(z)}) \\ &\leq \log \sum_{i=1}^{K} \hat{p}(z) \cdot \frac{1}{\hat{p}(z)} \\ &\leq \log(K). \end{aligned} \tag{A.4}$$

A.2 Behaviour of mIoU

A setting with three equally represented classes are presented here with varying fractions of TP:FP:FN. Three scenarios are created where the amount of sparsification is kept constant but the class chosen to perform the sparsification on is varied. The outcomes are presented in Table (A.1). It provides proof that the mIoU is sensitive to the sparsified-class(es).

Case	Class	GT	TP	FP	FN	Error: FP+FN	IoU		Error Ratio: (FP+FN)/TP
Original	1	300	100	100	100	200	33.33		2
	2	300	50	200	50	250	16.67		5
	3	300	250	25	25	50	83.33		0.2
							44.44	mIoU	
50 pixels removed from class 1 for the oracle									
1	1	300	100	50	100	150	40		1.5
	2	300	50	200	50	250	16.67		5
	3	300	250	25	25	50	83.33		0.2
							46.67	mIoU	
50 pixels removed from class 2 for the oracle									
2	1	300	100	100	100	200	33.33		2
	2	300	50	150	50	200	20		4
	3	300	250	25	25	50	83.33		0.2
							45.56	mIoU	
50 pixels removed from class 3 for the oracle									
3	1	300	100	100	100	200	33.33		2
	2	300	50	200	50	250	16.67		5
	3	300	250	0	0	0	100		0
							50	mIoU	

Table A.1: This table helps one quantify how the class chosen for sparsification has a significant impact on the mIoU, regardless of the amount of sparsification in one iteration.

In Table A.1, when class 3 with the lowest error ratio is chosen for sparsification, it increases mIoU by the maximum amount. This is the reason behind why it is hypothesized that mIoU could be ill-posed to be used as a sorting measure in calculating AUSE.

A.3 Supplementary results - Study 1

The impact of defocus corresponding to Zernike coefficient ω_4 on the Cityscapes dataset, for all the optical quality metrics discussed in Section 4. While the plots are with respect to the normalized Shapley value, the unnormalized Shapley value has been recorded and highlighted alongside the plots.

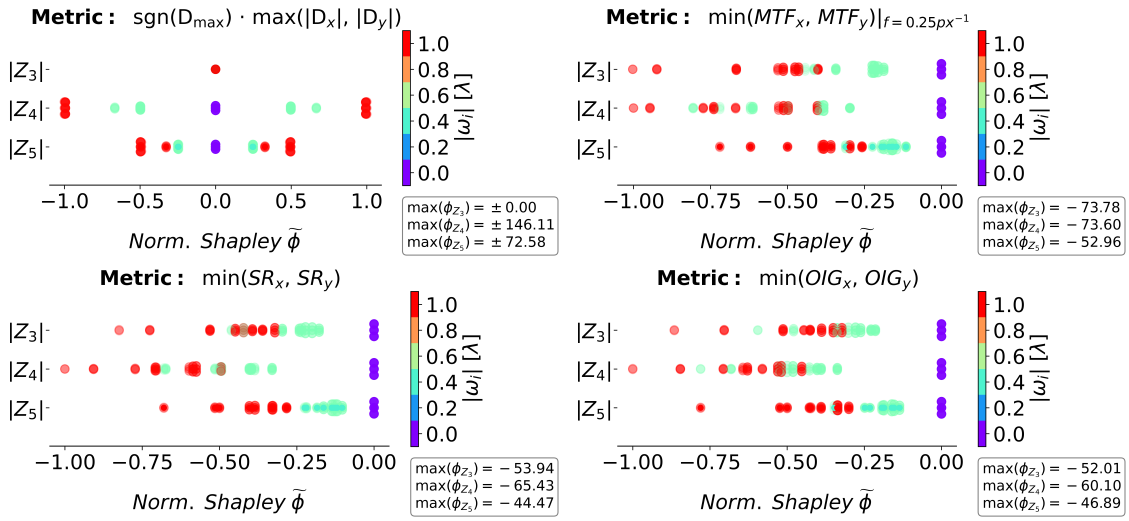

Figure A.1: The Shapley contributions for ω_3 and ω_4 for the case of MTF seem to be close in value for the study on Cityscapes. The negligibly small difference between the contribution of ω_3 and ω_4 in the case of MTF is neglected.

B Appendix

B.1 Noise-Image Method

The reverse-projection introduced by Koren et al. [34] and discussed in detail in Section 4.1.5 was implemented manually to study the difference in intensities between the raw ROI and the inverse-binned ROI, especially on the brighter side of the edge.

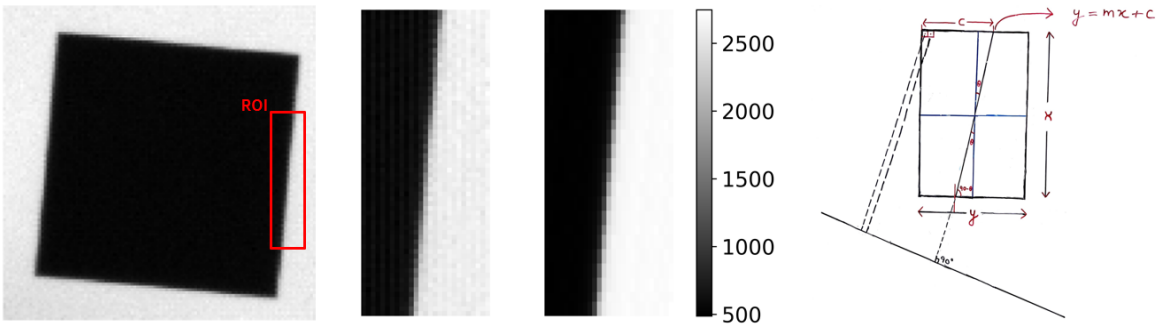

Figure B.1: The image on the left corresponds to a slanted square target from which individual edges are generally extracted for analysis. The slanted edge image on the left is directly extracted from the red region of interest (ROI). The image on the right is the filtered result of the reverse projection of the 4$\times$oversampled ESF (Edge Spread Function). The hand-drawn illustration is an attempt to visualize the projection of pixel intensities on the plane perpendicular to the slanted edge.

B.2 Algorithm - Information Capacity

The Noise-Image methodology (described in Section 4.1.5) consists of several steps starting from the noisy image of a slanted-edge target to the Information Capacity value C. The steps are sequentially described in Algorithm 1.

Algorithm 1: Noise-Image algorithm to calculate C, C_{max}

Input: Camera parameters, sensor parameters, slanted-edge target image(s) I, bit-depth of image(s).

Output: C_{NEQ}, C_{max}.

1 Calculate 4×oversampled edge profile $\mu_S(x)$ using ISO:12233 e-SFR algorithm.

2 Calculate the inverse-binned image I_{IB} by reverse-projecting the averaged $\mu_S(x)$ contents back to the original pixel coordinates.

3 Calculate noise image: I_N = I - I_{IB}.

4 Calculate the shifted 2D-FFT of noise image $F_{I_N}(u, v)$ with zero frequency components ($u = v = 0$) at the centre.

5 Perform radial averaging to calculate $NPS(f)$ with number of bins $N_{bins} \in [8, 32]$.

6 Calculate spatially dependent noise power from the Edge Variance (EV) method $\sigma_s^2(x)$ and subsequently apply correction factor: Normalize $NPS(f)$ using

$$\int NPS(f)df = \int \sigma_s^2(x)dx.$$

7 Calculate the Kernel function $K(f)$ by using

$$K(f) = \frac{MTF^2(f)}{NPS(f)}.$$

8 Calculate Noise Equivalent Quanta $NEQ(f)$ by using the normalized voltage ($V = \frac{V_{p-p}}{\sqrt{12}}$) where V_{p-p} is the peak-to-peak signal voltage

$$NEQ(f) = V^2 \times K(f).$$

9 Calculate Information Capacity C_{NEQ} using

$$C_{NEQ} = \int_0^{Ny} \log_2 (1 + NEQ(f))\, df.$$

10 Replace V_{p-p} by the maximum allowable value $V_{p-p-max} = 1$ to obtain maximum Information Capacity C_{max} using step 9. Detailed documentation in Koren et al. [34]

11 Repeat steps 1-10, for the entire batch of images.

References

[1] SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, 2016. https://www.sae.org/standards/content/j3016_202104/, Last accessed on 28-05-2024.

[2] Dimitris Milakis, Bart van Arem, and Bert van Wee. Policy and society related implications of automated driving: A review of literature and directions for future research. *Journal of Intelligent Transportation Systems*, 21(4):324–348, 2017.

[3] SAE International. 6 Levels of ADAS. https://www.sae.org/binaries/content/assets/cm/content/blog/sae-j3016-visual-chart_5.3.21.pdf, Last accessed: 28-05-2024.

[4] Fraunhofer IKS (Institute for Cognitive Systems). Technical requirements for autonomous driving, Blog Link, 01-01-2021. https://www.iks.fraunhofer.de/en/topics/autonomous-driving.html, Last accessed on 28-05-2024.

[5] Christian Krebs, Patrick Müller, and Alexander Braun. Impact of windshield optical aberrations on visual range camera based classification tasks performed by cnns. *London Imaging Meeting*, 2(1):83–83, 2021.

[6] Korbinian Weikl, Jeppe Revall Frisvad, Damien Schroeder, and Walter Stechele. Imaging through curved glass: Windshield optical impact on automotive cameras. In *ODS 2022: Industrial Optical Devices and Systems*, volume 12231, pages 55–64. SPIE, 2022.

[7] Dominik Werner Wolf, Markus Ulrich, and Alexander Braun. Windscreen optical quality for ai algorithms: Refractive power and mtf not sufficient. *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 5190–5197, 9 2023.

[8] Dominik Werner Wolf, Markus Ulrich, and Nikhil Kapoor. Sensitivity analysis of ai-based algorithms for autonomous driving on optical wavefront aberrations induced by the windshield. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4102–4111, 10 2023.

[9] ISO12233. Photography — Electronic still picture imaging — Resolution and spatial frequency responses. Standard, International Organization for Standardization, Geneva, CH, January 2017.

[10] Kenneth Parulski, Peter D Burns, Dietmar Wueller, and Hideaki Yoshida. Creation and evolution of iso12233, the international standard for measuring digital camera resolution. *Image*, 347:2, 2022.

[11] Jonathan B Phillips and Henrik Eliasson. *Camera image quality benchmarking*. John Wiley & Sons, 2018.

[12] Stephen Reichenbach, Stephen Park, and Ramkumar Narayanswarmy. Characterizing digital image acquisition devices. *Optical Engineering*, 30:170–177, 02 1991.

[13] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[14] Claude Elwood Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[15] Balint Mucsanyi, Michael Kirchhof, and Seong Joon Oh. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. 2024.

[16] Dominik Werner Wolf, Prasannavenkatesh Balaji, Alexander Braun, and Markus Ulrich. Decoupling of neural network calibration measures, 2024.

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[18] R Caruana. Predicting good probabilities with supervised learning. In *Proceedings of NIPS 2004 Workshop on Calibration and Probabilistic Prediction in Supervised Learning*, 2004.

[19] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[20] Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR, 2023.

[21] Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European Conference on Computer Vision*, pages 555–569. Springer, 2022.

[22] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[23] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[24] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

[25] Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

[26] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[27] Amiel Feinstein. A new basic theorem of information theory. 1954.

[28] M. J. Irland. Windshield optics. *Appl. Opt.*, 8(9):1787–1790, Sep 1969.

[29] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10193–10202, 2019.

[30] Frits Zernike. Diffraction theory of the knife-edge test and its improved form, the phase-contrast method. *Monthly Notices of the Royal Astronomical Society, Vol. 94, p. 377-384*, 94:377–384, 1934.

[31] Vasudevan Lakshminarayanan and Andre Fleck. Zernike polynomials: a guide. *Journal of Modern Optics*, 58(7):545–561, 2011.

[32] Dominik Werner Wolf, Markus Ulrich, and Alexander Braun. Novel developments of refractive power measurement techniques in the automotive world. *Metrologia*, 60, 9 2023.

[33] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005.

[34] Norman L Koren. Measuring camera information capacity with slanted-edges. *Electronic Imaging*, 35:454–1, 2023.

[35] On Semiconductor Corporation. ON Semi AR0820AT Sensor, Datasheet Request Link, 01-01-2021. `https://www.onsemi.com/products/sensors/image-sensors/AR0820AT#technical-documentation`, Last accessed on 28-05-2024.

[36] Ian A Cunningham and Rodney Shaw. Signal-to-noise optimization of medical imaging systems. *JOSA A*, 16(3):621–632, 1999.

[37] Brian W Keelan. Imaging applications of noise equivalent quanta. *Electronic Imaging*, 28:1–7, 2016.

[38] Oliver van Zwanenberg, Sophie Triantaphillidou, Robin B Jenkin, and Alexandra Psarrou. Estimation of iso12233 edge spatial frequency response from natural scene derived step-edge data. *Journal of Imaging Science and Technology*, 65(6):60402–1, 2018.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[40] Kira Maag, Matthias Rottmann, and Hanno Gottschalk. Time-dynamic estimates of the reliability of deep semantic segmentation networks. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 502–509. IEEE, 2020.

[41] Matthias Rottmann, Pascal Colling, Thomas Paul Hack, Robin Chan, Fabian Hüger, Peter Schlicht, and Hanno Gottschalk. Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[42] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[43] Matthias Rottmann and Marius Schubert. Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[44] Max-Heinrich Laves, Sontje Ihler, Karl-Philipp Kortmann, and Tobias Ortmaier. Well-calibrated model uncertainty with temperature scaling for dropout variational inference. *arXiv preprint arXiv:1909.13550*, 2019.

[45] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Feb. 2015.

[46] Kira Wursthorn, Markus Hillemann, and Markus Ulrich. Uncertainty quantification with deep ensembles for 6d object pose estimation. *arXiv preprint arXiv:2403.07741*, 2024.

[47] Mariella Dreissig, Florian Piewak, and Joschka Boedecker. On the calibration of underrepresented classes in lidar-based semantic segmentation, 2022.

[48] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[49] Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[50] Mariella Dreissig, Florian Piewak, and Joschka Boedecker. On the calibration of uncertainty estimation in lidar-based semantic segmentation. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4798–4805. IEEE, 2023.

[51] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[52] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *Inter. Journal of Robotics Research (IJRR)*, 2013.

[53] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[54] Thomas Mensink, Jasper Uijlings, Alina Kuznetsova, Michael Gygli, and Vittorio Ferrari. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9298–9314, 2021.

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. pages 234–241, 2015.

[56] Dominik Werner Wolf. Private Conversation, 2024.

[57] Eyke Huellermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 2021.

[58] Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952.

[59] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General pitfalls of model-agnostic interpretation methods for machine learning models, 2021.

[60] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

[61] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020.

[62] Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022.

[63] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR, 2019.

[64] Cornelius Emde, Francesco Pinto, Thomas Lukasiewicz, Philip HS Torr, and Adel Bibi. Towards certification of uncertainty calibration under adversarial attacks. *arXiv preprint arXiv:2405.13922*, 2024.

[65] Dominik Werner Wolf. Private Conversation, 2024.