DGK Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C | Dissertationen | Heft Nr. 878

Max Mehltretter

Uncertainty Estimation for Dense Stereo Matching using Bayesian Deep Learning

München 2021

Bayerische Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5290-1

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 378, Hannover 2021

DGK Veröffentlichungen der DGK

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften

Reihe C Dissertationen Heft Nr. 878

Uncertainty Estimation for Dense Stereo Matching using Bayesian Deep Learning

Von der Fakultät für Bauingenieurwesen und Geodäsie
der Gottfried Wilhelm Leibniz Universität Hannover
zur Erlangung des Grades
Doktor-Ingenieur (Dr.-Ing.)
genehmigte Dissertation

von

Max Mehltretter, M.Sc.

München 2021

Bayerische Akademie der Wissenschaften

ISSN 0065-5325 ISBN 978-3-7696-5290-1

Diese Arbeit ist gleichzeitig veröffentlicht in:
Wissenschaftliche Arbeiten der Fachrichtung Geodäsie und Geoinformatik der Leibniz Universität Hannover
ISSN 0174-1454, Nr. 378, Hannover 2021

Adresse der DGK:

Ausschuss Geodäsie der Bayerischen Akademie der Wissenschaften (DGK)

Alfons-Goppel-Straße 11 ● D – 80 539 München

Telefon +49 – 89 – 23 031 1113 ● Telefax +49 – 89 – 23 031 - 1283 / - 1100

e-mail post@dgk.badw.de ● http://www.dgk.badw.de

Prüfungskommission:

Vorsitzender: Prof. Dr. Philipp Otto

Referent: Prof. Dr.-Ing. habil. Christian Heipke

Korreferenten: Prof. Dr. Konrad Schindler (ETH Zürich)
Dr.-Ing. habil. Hamza Alkhatib

Tag der mündlichen Prüfung: 29.09.2021

ISSN 0065-5325 ISBN 978-3-7696-5290-1

Abstract

Motivated by the need to identify erroneous depth estimates, various methods for the estimation of uncertainty in the context of dense stereo matching have been presented in the literature in recent years. Especially, the introduction of deep learning-based methods and the accompanying significant improvement in accuracy have greatly increased the popularity of uncertainty estimation for dense stereo matching. Despite this remarkable development, most of the methods presented exclusively address the estimation of the uncertainty that arises from the data, called aleatoric uncertainty, and focus on the employed functional model only. In contrast, the epistemic uncertainty, which is embedded in the employed model and commonly caused by simplifications or incorrect model assumptions, is often neglected. However, to accurately quantify the uncertainty inherent in a process, it is necessary to consider all potential sources of uncertainty and to model their stochastic behaviour appropriately.

To approach this objective, the estimation of both, aleatoric and epistemic uncertainty in the context of dense stereo matching is addressed in this thesis. For the purpose of aleatoric uncertainty estimation, a novel Convolutional Neural Network architecture is presented that is trained with different stochastic models that are developed in the context of this work and that follow the concept of Bayesian deep learning. The quantification of epistemic uncertainty, on the other hand, is realised using a Bayesian Neural Network trained with variational inference. Combining both approaches, in this thesis a new holistic method to jointly estimate disparity and uncertainty is presented, taking into account both aleatoric and epistemic uncertainty.

To evaluate the performance of the method proposed and to investigate strengths and limitations, extensive experiments are carried out on four datasets considering real-world indoor and outdoor scenes. The results of these experiments demonstrate that both the functional and the stochastic models used to estimate aleatoric uncertainty outperform state-of-the-art methods in almost all scenarios. Moreover, it can be shown that the usage of a Bayesian Neural Network instead of a deterministic variant not only allows for epistemic uncertainty estimation, but also supports the task of dense stereo matching itself, reducing the amount of errors contained in the disparity maps obtained. Finally, the evaluation reveals the importance of considering both, aleatoric and epistemic uncertainty in order to achieve an accurate estimation of the overall uncertainty related to a depth estimate.

Keywords: Dense Stereo Matching, Depth Reconstruction, Uncertainty Quantification, Deep Learning, Bayesian Neural Network

Kurzfassung

Motiviert durch die Notwendigkeit fehlerhafte Tiefenschätzungen identifizieren zu können, wurden in den letzten Jahren verschiedene Methoden zur Abschätzung der Unsicherheit im Kontext der dichten Bildzuordnung vorgestellt. Insbesondere das Aufkommen von Methoden basierend auf dem Konzept des tiefen Lernens und der damit einhergehenden signifikanten Verbesserund der Genauigkeit, haben zur Steigerung der Popularität dieses Forschungsfeldes beigetragen. Von dieser bemerkenswerten Entwicklung abgesehen, beschäftigen sich die meisten dieser Methoden jedoch ausschließlich mit der Unsicherheit, die sich aus den Daten ergibt, auch aleatorische Unsicherheit genannt, und fokussieren sich dabei allein auf das funktionale Modell. Die epistemische Unsicherheit, die der verwendeten Bildzuordnungsmethode innewohnt und sich aus Vereinfachungen und inkorrekten Modellannahmen ergibt, wird hingegen häufig vernachlässigt. Um die Unsicherheit eines Prozesses jedoch akkurat bestimmten zu können, ist es notwendig, alle potenziellen Unsicherheitsquellen zu berücksichtigen und deren stochastisches Verhalten korrekt zu modellieren.

Auf dem Weg zu diesem Ziel, befasst sich diese Thesis mit der Abschätzung beider Arten von Unsicherheit im Rahmen der dichten Bildzuordnung. Zu diesem Zweck wird eine neue Convolutional Neural Network Architektur sowie verschiedene stochastische Modellen, die dem Konzept des Bayes'schen tiefen Lernens entsprechen, zur Abschätzung der aleatorischen Unsicherheit vorgeschlagen. Die Quantifizierung der epistemischen Unsicherheit wird hingegen über ein Bayes'sches Neuronales Netz realisiert, welches mittels Variationsinferenz trainiert wird. Mit der Kombination dieser beiden Ansätze, wird in dieser Thesis eine ganzheitliche Methode zur gemeinsamen Abschätzung von Disparität und zugehöriger Unsicherheit vorgestellt, in welcher sowohl aleatorische wie auch epistemische Unsicherheiten berücksichtigt werden.

Zur Evaluierung der vorgeschlagenen Methodik und zur Untersuchung ihrer Stärken und Limitierungen, wird eine Reihe an Experimenten auf vier verschiedenen Datensätzen durchgeführt, wobei sowohl Innen- wie auch Außenszenen berücksichtigt werden. Die Ergebnisse dieser Experimente belegen, dass sowohl das vorgeschlagene funktionale wie auch die stochastischen Modelle zur Abschätzung der aleatorischen Unsicherheit mit dem Stand der Technik mithalten können und diesen in vielen Szenarien sogar übertreffen. Zudem kann gezeigt werden, dass die Verwendung eines Bayes'schen Neuronalen Netzes neben der Abschätzung der epistemischen Unsicherheit auch zur Verbesserung der Genauigkeit der bestimmten Disparitätskarten beiträgt. Abschließend unterstreicht die Evaluierung die Bedeutung der Berücksichtigung beider Arten von Unsicherheit, um die Gesamtunsicherheit einer Tiefenschätzung akkurat bestimmen zu können.

Keywords: Dichte Bildzuordnung, Tiefenrekonstruktion, Unsicherheitsquantifizierung, Tiefes Lernen, Bayes'sches Neuronales Netz

Nomenclature

Abbreviations

BN	Batch Normalisation
BNN	Bayesian Neural Network
CCNN	Confidence Convolutional Neural Network
CNN	Convolutional Neural Network
CVA-Net	Cost Volume Analysis Network
ELBO	Evidence Lower Bound
GC-Net	Geometry and Context Network
KL	Kullback-Leibler
LFN	Late Fusion Network
LGC-Net	Local-Global Confidence Network
MAE	Mean Absolute Error
MCMC	Markov Chain Monte Carlo
MLP	Multilayer Perceptron
PER	Pixel Error Rate
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
SGM	Semi-global matching
VI	Variational Inference

Symbols

$KL(\cdot\|\cdot)$	Kullback-Leibler divergence of two probability distributions
$\hat{d}$	Single ground truth disparity value
$\mathbf{p}$	Pixel in a digital image, with $\mathbf{p} = (x, y)$ indicating its location as column and row
$\mathcal{A}_\mathbf{p}$	Local neighbourhood around a pixel $\mathbf{p}$
$\mathcal{D}$	Set of training or test data (containing pixels with known reference disparity only)
$\mathcal{L}$	Loss function
ϕ	Unknown parameters of a variational distribution
$\mathbf{C}$	Cost volume (corresponding to the left image of a stereo pair)

$\mathbf{D}$	Disparity map (corresponding to the left image of a stereo pair)
$\mathbf{I}_x$	Image with $x \in \{L, R\}$ referring to the left or right image of a stereo pair
$\mathbf{U}_x$	Uncertainty map with $x \in \{A, E\}$ referring to aleatoric or epistemic uncertainty, where $\mathbf{U}_x$ is expressed as a variance σ_x^2 per pixel
θ	Unknown parameters of a neural network
d	Single disparity estimate
f	Function representing the dense stereo matching process, with $f^{(x)}$, $x \in \{a, c, d\}$ being individual stages of this process
k	Index of a Monte Carlo sample

Contents

1 Introduction

The reconstruction of depth information from a stereoscopic image pair is a classical task in photogrammetry as well as in computer vision, while two such images also represent the minimum input for the well-known structure from motion problem. The latter refers to the concept of recovering 3D structures from the projected 2D motion field of a scene acquired with a moving sensor. A special case of this task is dense stereo matching, in which depth information is determined for every or at least a large majority of pixels within an image pair. Even after decades of intensive research, research questions on the topic of dense stereo matching remain open or newly arise due to novel technical possibilities, characterising this topic as a highly active field of research. The relevance and great interest in this topic is further demonstrated by the broad spectrum of its current deployment: Applications from the domains of robotics, the automotive industry, the medical sector as well as the creation of digital environment and surface models are often built on top of depth information obtained by dense stereo matching.

Independent of this success story and the often convincing results, image-based depth reconstruction remains to be a challenging task. The reason for its difficulty can be illustrated by interpreting depth reconstruction as the inverse operation to a perspective projection: Since projecting a 3D scene to a 2D image plane results in a dimensionality reduction, the inverse operation does not have a unique solution in general. To determine a solution nevertheless, the identification of point correspondences between the two images of a stereo pair is typically a prerequisite. Whereas humans are able to estimate the depth of objects with the greatest of ease in various situations, particularly demanding conditions, such as poor illumination, low-textured or non-Lambertian surfaces as well as occlusions, still pose a great challenge to related algorithms. The ambiguities that arise from such conditions may prevent dense stereo matching approaches from identifying the correct correspondences for all pixels and thus characterises stereo image-based depth reconstruction as an ill-posed task. Most approaches presented in the literature, however, focus on maximising the accuracy for a specific application and data domain, while neglecting the problem of potentially incorrect matching results by treating all depth estimates as equally reliable. However, it is often possible to assess how trustworthy the estimated depth information is. This task becomes crucial for safety-critical applications: Assigning a high level of uncertainty to erroneous predictions can prevent a system from taking wrong decisions with potentially fatal consequences.

As in many other fields, the introduction of deep learning into dense stereo matching brought a significant improvement in accuracy. Nevertheless, this type of approach is in general very sensitive to all kinds of variations contained in the data to be processed (Tonioni et al., 2017), as a specific characteristic can typically only be handled if it has been learned during training. A situation not observed during this phase, on the other hand, will lead to a poor matching accuracy with

high probability. This implies that assumptions sometimes explicitly formulated for hand-crafted matching approaches, are implicitly introduced to learned approaches by the selection of training data also. Due to the high range of variations within the real world, however, it is probably impossible to provide divers enough training data to learn all relevant characteristics necessary to reliably operate in a real world scenario (Zendel et al., 2017). Thus, the use of learning-based techniques leads to an aggravation of the tension between accuracy and reliability in real world settings, which already became evident in the context of conventional approaches. Consequently, the quantification of uncertainty remains an important topic, especially in the context of deep learning-based stereo matching.

1.1 Problem Statement

To accurately quantify the uncertainty inherent in a process, it is necessary to consider all potential sources of uncertainty. In this context, two types of uncertainties are generally distinguished, following the taxonomy introduced by Hacking (1975): aleatoric and epistemic uncertainty. Aleatoric uncertainty is contained in the data and caused by variable, non-deterministic or simply unpredictable behaviour of a process under consideration. In contrast, epistemic uncertainty accounts for incorrect or inaccurate model hypotheses. Interpreting this differentiation in a Bayesian way, aleatoric uncertainty can be modelled as a probability distribution over the predictions, whereas epistemic uncertainty refers to a probability distribution over the parameters of a model.

To address the task of aleatoric uncertainty estimation for dense stereo matching, several hand-crafted as well as deep learning-based approaches have been presented in the literature already. However, most of them do not interpret this task in a Bayesian way, resulting in a unit-less measure of reliability, typically between zero and one. Such results are not only hard to interpret and of limited expressiveness for applications built on top of depth information obtained via dense stereo matching, but it is also not trivial to combine them with uncertainties originating from other sources. In contrast, the quantification of epistemic uncertainty is rarely discussed in the literature with respect to more complex tasks from the field of photogrammetry, which is particularly true for dense stereo matching. The proper parametrisation of a complex model is already a challenging task commonly approached empirically. Consequently, it is even more difficult to not only estimate a point value characterising a parameter, but a probability distribution. However, recent developments in the field of Bayesian deep learning, and Bayesian Neural Networks (BNNs) in particular, promise to allow learning this information from training data.

In this thesis, the task of uncertainty quantification in the context of dense stereo matching is approached from a Bayesian perspective. For this purpose, different techniques from the domain of Bayesian deep learning are adapted and newly developed to jointly estimate aleatoric and epistemic uncertainty. In this context, research questions related to the following aspects are investigated:

- Which information is valuable for the prediction of the aleatoric uncertainty associated to a disparity estimate? How can useful features be extracted from this information and what assumptions need to be made in this context?

- How is the uncertainty of disparity estimates distributed and which conclusions can be drawn from this distribution with respect to the formulation of a reasonable uncertainty model? Which specific characteristics need to be considered and which simplifications might be acceptable?

- Which opportunities exist to quantify epistemic uncertainty, and are BNNs a reasonable choice considering theoretical as well as practical aspects? What are the consequences of transforming a Convolutional Neural Network (CNN) from a deterministic to a probabilistic formulation? Which assumptions and simplifications need to be made in order to be able to train such a network? How are the quality of the depth estimates and the predicted uncertainties affected and are there any trade-offs between the two to be investigated or balanced?

1.2 Contributions

To address the previously described problem, a new holistic approach for the task of uncertainty quantification in the context of dense stereo matching is presented in this work. For this purpose, aleatoric as well as epistemic uncertainty are modelled in a Bayesian deep learning framework, resulting in the following main contributions:

- Design of a novel CNN architecture able to predict different measures of uncertainty, such as confidence scores or standard deviations under the assumption of a certain probability distribution. This network operates on cost volumes, which promise to contain more information than disparity maps, which are extracts of such volumes considering only disparity values with minimal cost. Moreover, the presented approach is able to process cost volumes originating from various dense stereo matching methods and allows to estimate the associated uncertainty either subsequent to the depth reconstruction or as part of it, estimating depth and uncertainty jointly.

- Development of two novel probabilistic mixture models to quantify aleatoric uncertainty in the context of dense stereo matching. Building on a Laplacian distribution used as a baseline, these models consider the ability to locate the true disparity with respect to the characteristics of the observed scene or the influence of random noise, respectively. Thus, the proposed models overcome the limitations of widely used confidence-based approaches that are unable to quantify the uncertainty in pixels or metric units, as well as those of the uni-modal baseline which often oversimplifies the actual error distribution by stating strong assumptions that are only valid for some of the disparity estimates.

- Realisation of a BNN, which allows to jointly estimate depth and epistemic uncertainty. For this purpose, an end-to-end trainable CNN architecture is adapted, which has already proven to be well-suited for the task of dense stereo matching. More precisely, probabilistic convolutional layers are incorporated, which are trained using Variational Inference (VI).

- Combination of the concepts proposed for the estimation of aleatoric and epistemic uncertainty. The resulting approach allows to estimate depth as well as aleatoric and epistemic uncertainty jointly.

1.3 Thesis Outline

In the remainder of this thesis, first an overview of relevant fundamentals on the topics of dense stereo matching, uncertainty quantification and deep learning are provided in Chapter 2. In this context, special emphasis is put on common challenges that cause uncertainty in image-based depth reconstruction processes as well as on the Bayesian deep learning framework which serves as basis for the proposed methodology. In Chapter 3, relevant literature is reviewed and discussed considering both, works that focus on the task of dense stereo matching itself as well as those that investigate uncertainty quantification related to this task. A new method for uncertainty estimation in the context of dense stereo matching is presented in Chapter 4, elaborating on individual approaches for the quantification of aleatoric (Sec. 4.2) and epistemic uncertainty (Sec. 4.3) first, before fusing them into a joint estimation procedure (Sec. 4.4). In Chapter 5, the experimental setup of this work is introduced, including a review of the datasets used, an overview of the training procedures of the presented deep learning-based approaches as well as a discussion of the metrics used to evaluate the results, which are presented and discussed in Chapter 6. Finally, this thesis closes with conclusions and an outlook on potential future works in Chapter 7.

2 Basics

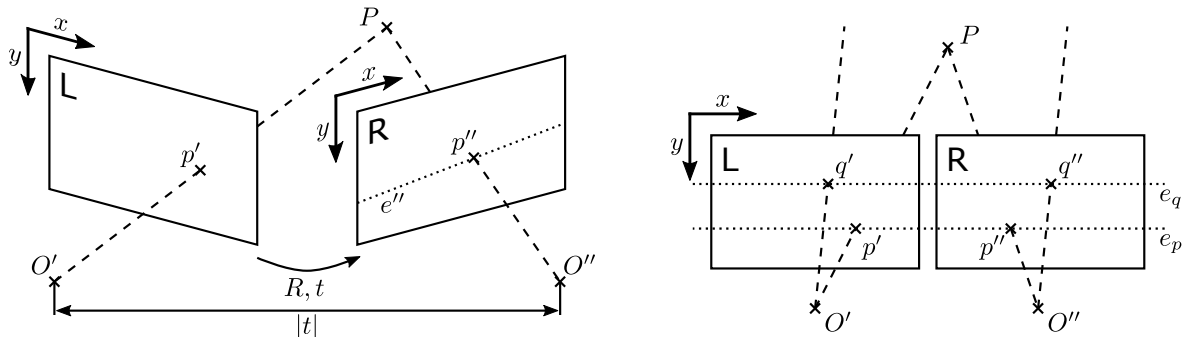
The concepts presented in this thesis correspond to the field of uncertainty quantification in the context of dense stereo matching and are based on techniques from the deep learning domain. Thus, in this chapter, basic definitions and a taxonomy commonly used to examine dense stereo matching approaches are reviewed first, see Section 2.1. In Section 2.2, the field of uncertainty quantification is briefly introduced, focusing on the differentiation into aleatoric and epistemic uncertainty and their interpretation in the context of dense stereo matching. Finally, basic concepts of deep learning are reviewed in Section 2.3. Due to the application domain relevant for the present work, a more detailed discussion is given on the idea of CNNs, which are commonly employed in the context of image-related tasks, as well as the idea of BNNs, which inter alia allow to estimate epistemic uncertainty.

2.1 Dense Stereo Matching

The objective of dense stereo matching is the determination of correspondences for all pixels of a specified reference image in one or multiple other images that show the same scene from slightly different viewpoints. With respect to the term *stereo*, derived from the Greek word $\sigma\tau\epsilon\rho\epsilon o\zeta$ (stereos), such correspondences describe a spatial relation in the sense that they depict the same point in 3D object space. Thus, with the identification of matching image points, stereo matching allows to reconstruct depth information via triangulation, that was lost due to the dimensionality reduction taking place as part of the imaging process.

2.1.1 Terminology and Practical Simplifications

In the general case, in which no information regarding the spatial alignment of the images is provided, all pixels must be considered when searching for correspondences. However, since this procedure is computational intensive and prone to errors caused by ambiguities, in the context of this work, the intrinsic orientations of all images as well as the relative orientations between the images is assumed to be known. Introducing this assumption, the search space can be reduced from the entire image to an epipolar line, thus simplifying the matching task from 2D to 1D (cf. Fig. 2.1a). Additionally, only the two image case (a reference image plus one matching image) is considered and it is assumed that the view point deviation between these images mainly consists of a shift in the direction of the x axis of the reference image. For simplicity, these images will be referred to



(a) Epipolar geometry (b) Planar rectification

Figure 2.1: **Simplification of the matching process with known stereo calibration.** (a) The relation of a stereo image pair with known intrinsic and relative orientation can be described via epipolar geometry, where O', O'' are the projection centres and R, t are the rotation and translation between these centres. If two points p' and p'' depict the same object point P in the left and right image, respectively, p'' is located on the epipolar line e'' corresponding to p', which reduces the complexity of the matching task from 2D to 1D. (b) A planar rectification further simplifies the matching procedure, because all epipolar lines e are parallel and horizontal so that corresponding points are located in the same image row in both images.

as left and right image in the remainder of this work, with the left image being considered as the reference image.

The combination of the previously introduced assumptions allows the application of a planar rectification which projects both images to a common image plane. As a result of this rectification, the optical axes of both cameras are parallel and all epipolar lines are horizontally aligned. In consequence, corresponding pixels are located in the same image row in both images, as illustrated in Figure 2.1b, which further simplifies the matching process from a practical perspective. The deviation between the image columns x of corresponding pixels in the left and right image, is referred to as disparity or parallax d and can be described as $d = x_L - x_R$. Such disparity allows to draw conclusions regarding the distance between the common image plane of the reference and the matching image and the corresponding 3D point in the model coordinate system. While a disparity of zero implies parallel viewing rays and thus an infinite distance, the higher the disparity, the closer the associated three-dimensional object point is located to the common image plane (see Fig. 2.1b). In mathematical terms, an inverse proportional relationship exists between disparity d and distance Z in the model coordinate system:

$$Z = c\frac{|t|}{d}\,, \tag{2.1}$$

where c is the common principal distance of both images after rectification and $|t|$ is the base length between the two projection centres.

2.1.2 Taxonomy of the Matching Process

As described in the previous section, the determination of disparity, and therefore also of depth, relies on the identification of corresponding image points. On the basis of a planar rectified stereo

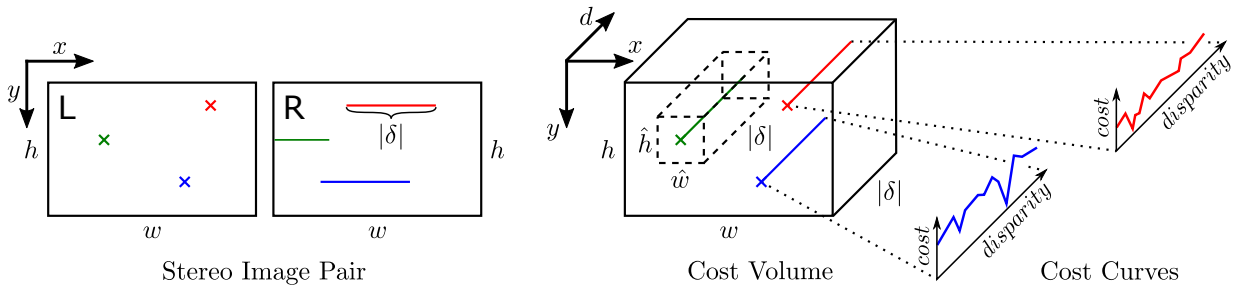


Figure 2.2: **Relation between a stereo image pair, a cost volume and cost curves.** If not otherwise specified, a cost volume refers to the left image of a planar rectified stereo image pair. In x and y direction it has the same width w and height h as the images, while the size in d direction is defined by the considered disparity search range δ. Every extract of this cost volume (shown with dashed lines) has the same size in d direction as the whole volume, independent of its width $\hat{w}$ and height $\hat{h}$. The cost volume consists of one cost curve per pixel in the reference image, encoding the matching costs of potential correspondences in the right image along the epipolar line and within the specified disparity range δ. Source: Mehltretter and Heipke (2021).

image pair, a variety of different approaches have been proposed for this purpose over the last decades in the literature. Using the taxonomy introduced by Scharstein and Szeliski (2002), however, the fundamental concept underlying most of these approaches can be examined and discussed in a uniform way. For this purpose, the taxonomy differs between four different steps, of which a subset is realised in most approaches presented in the literature: cost computation, cost aggregation, disparity computation and optimisation, and disparity refinement.

In the first step, the cost computation, the dissimilarity for potential correspondences is determined. For this purpose, a pixel from the left image $\mathbf{I}_L$ is compared against all pixels located on the corresponding epipolar line in the right image $\mathbf{I}_R$ within a search range $\delta = [d_{min}, d_{max}]$. The result is a cost curve encoding the dissimilarity obtained per pixel, also referred to as matching cost, with respect to the disparity. The individual curves are typically summarised in a three-dimensional cost volume $\mathbf{C}$ (sometimes also called disparity space image), a concept originally proposed by Yang et al. (1993) and Intille and Bobick (1994):

$$\mathbf{C}(x, y, d) = f^{(c)}(\mathbf{I}_L(x, y), \mathbf{I}_R(x - d, y)) \,, \tag{2.2}$$

where $f^{(c)}$ refers to the cost function used to measure the dissimilarity. Such a cost function is often defined based on the assumption of photoconsistency, e.g., computing the sum of absolute or squared differences, under the expectation that matching pixels were assigned the same grey scale or RGB value. However, also other statistics defined on the images can be used to define a dissimilarity metric, such as the normalised cross-correlation coefficient, least squares matching (Förstner, 1982), Census (Zabih and Woodfill, 1994) or mutual information. The relation between a stereo image pair, a cost volume and its cost curves is illustrated in Figure 2.2.

The subsequent cost aggregation step aims to reduce noise and local outliers inherent in the costs resulting from the first step. For this purpose, a weighted aggregation scheme w is applied to smooth the initial cost volume $\mathbf{C}$ along the spatial dimensions x and y:

$$\mathbf{C}'(\mathbf{p}, d) = w(\mathbf{p}) \cdot \mathbf{C}(\mathbf{p}, d) + \sum_{\mathbf{q} \in \mathcal{A}_\mathbf{p}} w(\mathbf{q}) \cdot \mathbf{C}(\mathbf{q}, d) \,, \tag{2.3}$$

where $\mathbf{p} = (x, y)$ is a pixel in the reference image and $\mathcal{A}_\mathbf{p}$ is its associated local neighbourhood. This procedure implies that all pixels within the local neighbourhood have similar cost distributions and thus possess a similar disparity. The simplest variant of such an aggregation scheme is a box filter, a quadratic filter mask computing an equally weighted average. However, also more sophisticated approaches exist which consider local image characteristics to determine individual filter shapes and weights per pixel $\mathbf{p}$, such as intensity differences or spatial distances between pixels in the reference image. Considering such additional information in the definition of the aggregation scheme, undesired smoothing can be reduced, for example, at depth discontinuities. It is worth noting that the cost aggregation step is not present in all dense stereo matching methods, but is particularly important for local approaches, which do not perform a global cost optimisation.

The objective of the third step of the taxonomy is the extraction of a globally optimal disparity map $\mathbf{D}$ from a (smoothed) cost volume $\mathbf{C}$. This task is often interpreted as an energy minimisation problem, considering the previously computed costs together with specific model assumptions:

$$E(\mathbf{D}) = \sum_{\mathbf{p}} \left(\mathbf{C}(\mathbf{p}, \mathbf{D}(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{A}_\mathbf{p}} g(\mathbf{D}(\mathbf{p}), \mathbf{D}(\mathbf{q})) \right) , \tag{2.4}$$

where g is a function that penalises deviations from the model assumptions in a local neighbourhood. A typical example for such an assumption is piece-wise smoothness of the scene geometry. Since the determination of a disparity map with minimum energy is a NP-hard problem (Boykov et al., 1999), Equation 2.4 is only solved approximately, e.g., via methods based on Graph Cuts or Belief Propagation (Tappen and Freeman, 2003).

In the last step of the taxonomy, a determined disparity map is refined by applying post-processing techniques. The repertoire of such techniques is rather broad, ranging from simple filter operations, such as the use of a median filter for noise reduction, to more sophisticated approaches, for example, used to eliminate erroneous disparity estimates. The refinement step also includes approaches used to determine sub-pixel accurate disparities, whereas disparity maps extracted from the cost volume directly are typically limited to integer disparities because of the discrete nature of cost volumes.

2.1.3 Challenges and Common Assumptions

As stated in the introduction, the reconstruction of depth information from a stereo image pair is an ill-posed task without a unique solution in general. Multiple scenarios exist, such as occlusion or texture-less areas, where the comparison of individual pixels or image patches delivers no or several potential correspondences that are equally similar with respect to a specified metric. To resolve these ambiguities, additional assumptions are often introduced, constraining the matching process. The most common and for this work most relevant assumptions are discussed in the following together with common problems violating these assumptions.

Photoconsistency

Photoconsistency implies that a depicted scene has the same appearance in both images of a stereo pair and is one of the most commonly applied assumptions. Under this assumption, intensities or colour values of pixels can be compared directly without the need of normalising them or applying matching on a more abstract level. Thus, the matching task is simplified significantly. However, the presence of non-Lambertian surfaces, vignetting, noise or varying illumination as well as the combination of different imaging modalities, such as RGB and thermal, typically violate this assumption. Consequently, photoconsistency is typically only expected in an attenuated form, for example, via the usage of matching cost metrics that are robust against certain types of noise or global differences in illumination and contrast (Zabih and Woodfill, 1994). Alternatively, the images can also be transformed into a representation that is invariant towards such influences before computing the matching costs (Mouats et al., 2015), allowing to tolerate deviations from the expectation of photoconsistency to a certain extent.

Unique Matching

Based on the concept of epipolar geometry, a unique assignment exists between a 3D object point and its 2D depictions in image space. Based on this assumption, most stereo matching methods derive the existence of a unique one-to-one assignment between all points in both images of a stereo pair. This procedure implies that every point that is visible in one image also exists in the other image and requires that such correspondences can be identified unambiguously. In practise, both assumptions are addressed implicitly by assigning every image point the disparity, and thus the point in the other image, that results in the lowest matching cost (optionally, considering additional optimisation objectives such as piece-wise smooth surfaces). Violations of these assumptions are often only considered in a post-processing step, such as a left-right consistency check, and filtering out disparities that assign several image points from one image to a single point in the other image. However, the formulated assumptions are violated rather often and because of several reasons. On the one hand, multiple potential matches may exist. This can be caused by texture-less regions or repetitive patterns, where it is not possible to unambiguously decide which image point is the correct match. Another potential reason is the existence of reflections and semitransparent surfaces, where multiple 3D object points are projected to the same image point. On the other hand, also image points without a correspondence in the other image exist. This scenario is typically caused by occlusions or because the projection of the 3D object point is located outside of the visible extract of the image plane.

Piece-wise Smooth Surfaces

Most dense stereo matching methods constrain the matching process by the assumption of piece-wise smooth surfaces. The expectation is that depth changes smoothly between the majority of adjacent pixels and that depth discontinuities appear comparably seldom (Marr and Poggio, 1979). Thus, potential matches implying a depth that deviates from the expected smooth surface are

penalised in the optimisation process or are directly discarded. This assumption may be introduced via penalties for discontinuities between pixels (Hirschmuller, 2008), between image segments (Hong and Chen, 2004) or between primitives such as triangles (Bulatov et al., 2011). While segment-based approaches are relatively robust against noisy initial depth estimates, they generally tend to oversimplify the geometry of a scene. Primitive-based approaches, on the other hand, allow to preserve more details, but require a set of well-distributed image points with reliable depth estimates as prior information (Höllmann et al., 2020). A more sophisticated approach is the regularisation of depth using shape templates for objects within a scene that are selected based on prior semantic knowledge (Guney and Geiger, 2015).

Consistency of Image Gradients and Depth Discontinuities

It is frequently assumed that colour or intensity gradients in an image coincident with depth discontinuities, e.g., to decide where piece-wise smoothness is to be assumed and where abrupt depth changes should be tolerated (Hirschmuller, 2008). While this assumption often holds at object boundaries, such boundaries may not lead to image gradients if the colours or intensities of adjacent objects are similar. Also the opposite case, that an image gradient exists but it is not caused by a depth discontinuity, violates this assumption and is frequently caused by strong textures or shadows cast on the observed scene. With respect to the application example mentioned at the beginning, a violation of this assumption either leads to smooth depth transitions instead of sharp edges or to artificial depth discontinuities that do not exist in the real scene.

2.2 Uncertainty Quantification

Uncertainty quantification deals with the process of quantitatively describing and minimising different kinds of uncertainties inherent in a system for which certain aspects are not perfectly known. In this context, two types of uncertainties are often distinguished, following the taxonomy originally introduced by Hacking (1975): aleatoric and epistemic uncertainty. Although this classification has its origins in the field of philosophy, it is commonly applied to applications related to risk analysis (Van Asselt and Rotmans, 2002; Riesch, 2012), and more recently it has also been adapted to describe and analyse potential sources of uncertainty in the context of deep learning (Der Kiureghian and Ditlevsen, 2008; Kendall and Gal, 2017; Liu et al., 2019).

Following this taxonomy, aleatoric uncertainty, also referred to as stochastic uncertainty (Helton, 1994), random uncertainty (Henrion and Fischhoff, 1986) or irreducible uncertainty (Van Asselt and Rotmans, 2002), is typically caused by so called natural variability and are contained in the data or measurements. In this context, natural variability is understood as variable, non-deterministic or simply unpredictable behaviour of a process under consideration. Epistemic uncertainty, on the other hand, accounts for limited knowledge of a problem domain, either caused by an insufficient amount of observations or because the domain is too complex to be considered as a whole, and simplifications that are applied while defining a predictive model (Van Asselt and Rotmans, 2002). This category of uncertainty, also known as systemic (Henrion and Fischhoff, 1986) or reducible

uncertainty (Ditlevsen and Madsen, 1996), can be further subdivided into model form uncertainty, which reflects doubts regarding the structural correctness of the employed model, and parametric uncertainty, describing the uncertainty regarding the parametrisation of a model that is believed to be structurally correct (Sullivan, 2015). Moreover, differences between aleatoric and epistemic uncertainty can not only be seen in a theoretical context, but also in the practical approaches of their estimation: Aleatoric uncertainty can typically be predicted based on the data directly, epistemic uncertainty, on the other hand, is often estimated based on sampling approaches (Kendall and Gal, 2017).

From the perspective of dense stereo matching, aleatoric uncertainty accounts for effects such as sensor noise, occlusion and matching ambiguities caused, for example, by texture-less areas or repetitive patterns within a scene. Taking only into account the two image case, these effects cannot be minimised using more (diverse) training data, but are independent of the knowledge regarding the problem domain, which is illustrated by the following example: Even if the phenomenon of occlusion is perfectly modelled in a dense stereo matching method, a certain level of uncertainty regarding the depth of a depicted scene in occluded regions - the aleatoric uncertainty - remains because depth cannot be triangulated if an object point is visible only in a single image. Epistemic uncertainty, on the other hand, considers assumptions that simplify the matching process, such as those discussed in Section 2.1.3, and characteristics that are missing in the definition of this process (or in case of deep learning-based approaches in the training data), such as features and shades that imply a certain geometric shape. Consequently, a more diverse set of training data can reduce the amount of epistemic uncertainty embedded in a learned model, assuming that the structure and parametrisation of this model allow for such a reduction. Note that the assignment of specific effects to either aleatoric or epistemic uncertainty is not fixed, but depends on the definition of the problem domain. If, for example, the restriction to the two image case is relaxed to multi-view stereo, occlusion can be understood as a source of epistemic uncertainty, because taking additional images into account may reduce the uncertainty regarding the depth of object points that are occluded when considering only two images.

While the taxonomy of Hacking (1975) generally allows to investigate and model different sources of uncertainty independently, it may not consider interactions between them properly, resulting in an over-simplification of reality. Limited knowledge as a typical representative of epistemic uncertainty, for example, can result from natural variability (which is categorised as aleatoric uncertainty) if the ability to measure that variability is limited (Van Asselt and Rotmans, 2002). Thus, a clear distinction between different such sources is not necessarily possible in more complex scenarios (Riesch, 2012) and the assignment to either aleatoric or epistemic uncertainty may vary depending on the model employed and the context of its application (Der Kiureghian and Ditlevsen, 2008), as discussed earlier for the example of occlusion in the context of dense stereo matching. However, since these individual uncertainties are typically combined to describe particular events, the overall uncertainty present in a model can often nevertheless be approximated accurately.

2.3 Deep Learning

Deep learning and neural networks are concepts highly relevant for the approaches developed in the context of this thesis. Therefore, in this section, a general overview is given on this topic first, before focusing on the concepts of CNNs and BNNs. For a more general and detailed introduction into the field of neural networks and deep learning, the interested reader is referred to the standard literature such as the text books of Bishop (2006) and Goodfellow et al. (2016).

The concept of neural networks in combination with deep learning allows to model complex non-linear relations between observations and state estimations. For this purpose, the function $\mathcal{F}^*$ underlying this relation is approximated by a concatenation of multiple non-linear functions, realised as a network of neurons which are parameterised by their weights and biases θ:

$$y = \mathcal{F}^*(x) \approx \mathcal{F}_\theta(x)\,. \tag{2.5}$$

Every neuron computes its output as the weighted sum of its input information, offset by biases, before the result is processed with a non-linear function (also referred to as activation function), such as the Rectified Linear Unit (ReLU) (Glorot et al., 2011). Multiple such neurons operating in parallel form a so called layer, while multiple sequentially connected layers form a neural network. If a network consists of more than three cycle-free connected layers, it is commonly called Multilayer Perceptron (MLP) or feed forward neural network, whereas the number of layers is typically referred to as the depth of the network, which also coins the term deep learning. Every layer that is not the first or last layer of a network is called hidden layer. This term originates from the fact that such layers are not directly connected to the incoming or outgoing information of a network, making them invisible to operands outside of the network. The existence of such hidden layers allows a neural network to learn various representations of the input data on different levels of abstraction. Learning such representations is one of the reasons why neural networks often generalise well to unseen data that can be described by the same abstract representations that were learned from the training data. Moreover, this marks one of the major differences between deep learning-based and classical machine learning approaches. A specific variant of hidden layers are fully-connected ones, in which all nodes of this layer are connected to all nodes of the subsequent layer.

In general, the employment of neural networks can be distinguished into two phases: training and testing. During testing, the network parameters, namely the weights and biases, are kept constant and are simply used by the neurons to compute a result for given input data during a so called forward pass. To find suitable values for these parameters in the first place, they are iteratively adapted during training with respect to the training data and an optimisation objective. Typically starting with randomly initialised weights and biases set to zero, a prediction for a single or a batch of training samples is computed, exactly as it is done during testing. For this purpose, the initial values for the weights are commonly drawn from a zero-centred normal or uniform distribution with the width or limits of this distribution, for example, set based on the number of network nodes (Glorot and Bengio, 2010). The quality of this prediction is evaluated with respect to a specified optimisation objective, commonly referred to as loss function. In case of a supervised training procedure, which is applied in the present work, this evaluation is realised as comparison against a reference solution. Based on the concept of Stochastic Gradient Descent, e.g., in form of

ADAM (Kingma and Ba, 2015) or RMSprop (Tieleman and Hinton, 2012), the difference between prediction and reference is used to update the network parameters iteratively. The update itself is realised via backpropagation, in which the partial derivations of the neurons of a network are recursively computed with respect to the inputs of these neurons, aiming to reduce the difference between prediction and reference in the next iteration. The adaptation of the weights and biases in each neuron depends on the respective derivations and the learning rate, a hyper-parameter that specifies the amount of change that is applied during parameter update of a single training step. Repeating this alternation of forward pass and backpropagation many times, the network parameters are adapted to the information contained in the training data, allowing the network to predict results that converge against the reference solution.

A critical part of the training process is the loss function. It not only defines the optimisation objective, but also influences the convergence behaviour. Commonly, two major types of loss functions are distinguished with respect to the application domain of a neural network: classification and regression. In the case of classification, the network assigns probabilities to a predefined set of possible classes and the class with the highest probability is typically chosen as result. It is to be noted, that in this scenario the classes are typically unordered, no semantic relationships are considered nor is it possible to determine the distance between two classes. Often, variants of the cross entropy are used as loss functions for classification problems, but the hinge or logistic loss may also be employed (Rosasco et al., 2004). In a regression scenario, the network predicts real numbers in a interval specified in advance. Typical examples for regression-based loss functions are adaptations of the L1 and L2 norms, commonly referred to as Mean Absolute Error (MAE) and Mean Squared Error (MSE), respectively, if the error is averaged over multiple data samples.

2.3.1 Convolutional Neural Networks

The concept of CNNs extends the ideas of deep learning described so far by introducing specialised types of layers, of which convolutional layers are the most prominent. As the term CNN implies, a filter kernel a and an input signal b are convolved in order to compute an (intermediate) result $(b*a)$, regularly referred to as feature map, as illustrated in Figure 2.3. Note that in most cases multiple filter kernels are learned per layer instead of a single one. In this context, the depth of such a kernel is specified by the number of channels in the feature map to be processed, whereas the number of filter kernels in a layer specifies the number of channels in the resulting feature map. In contrast to the general concept of MLPs, where the incoming information of a node is commonly combined neglecting the topology, for example, using a weighted average, CNNs preserve the spatial structure of an input signal. This in turn requires that the incoming information is structured in a grid-like manner, which is, for example, the case for digital (single or multichannel) images. Moreover, using the same filter kernel over the whole image allows to detect a certain feature independent of its position, characterising convolutional operations as translation invariant. Finally, because the same filter kernel is applied to the whole input signal, this kernel has to be learned only once, resulting in spatially shared weights. Thus, the amount of parameters to be trained can be reduced drastically compared to the naive use of a basic MLP. Benefiting from these advantages, CNNs have been successfully employed for many different photogrammetric and computer vision tasks,

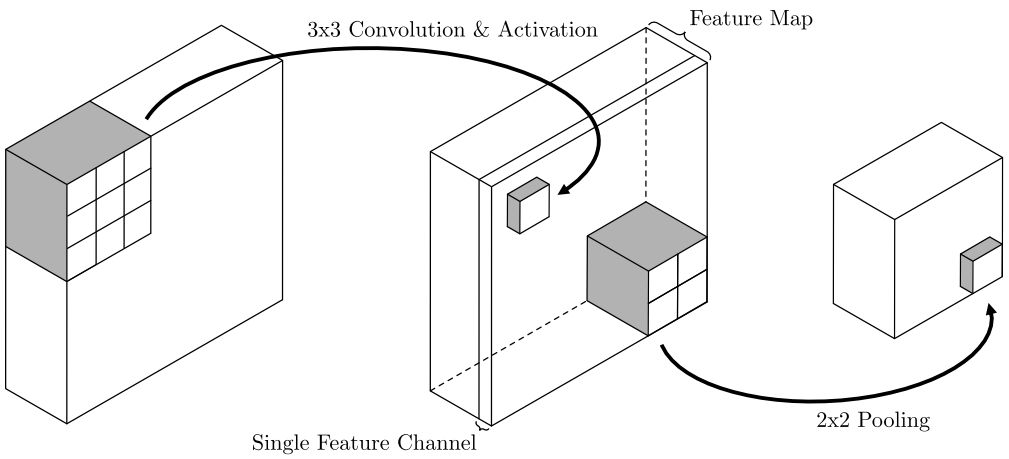


Figure 2.3: **Principle of convolutional and pooling operations.** In this exemplary setup, a 2D convolutional layer, consisting of multiple 3x3 filter kernels, is followed by an activation function to compute an intermediate feature map from a given input. Every individual filter kernel considers all channels of the input to compute one feature channel of the output. The channels resulting from multiple kernels are combined to form the intermediate feature map. In the second step, a 2x2 pooling operation is applied to downsample the intermediate feature map.

such as image classification (He et al., 2016), object detection (Girshick, 2015), pedestrian tracking (Leal-Taixé et al., 2016; Nguyen and Heipke, 2020), semantic segmentation (Marmanis et al., 2016), feature matching (Chen et al., 2020) and stereo depth estimation (Kendall et al., 2017b).

While convolutional layers often constitute the majority of layers and fully-connected ones are commonly used as final layers in CNNs designed for classification tasks, pooling operations are the third important type of layers in the context of CNNs. A pooling function typically operates on the output of a single or a sequence of convolutional layers (including a non-linear function and normalisation of the output), computing a summary statistic in a certain local context. The most prominent realisation of this principle is max pooling (Zhou and Chellappa, 1988), which extracts the maximum response of a local neighbourhood defined on a feature map. In the output of the pooling operation, only this value is kept to describe the feature map extract defined by the local neighbourhood it originates from, resulting in a reduction of the spatial dimension of the feature map. Thus, in contrast to strided convolutions, where the convolutional filter kernel is only applied on every n-th pixel with n specifying the stride, pooling does not result is a regular sampling of a feature map. However, both ways to downsample feature maps enlarge the receptive field and thus provide information from a larger context to subsequent layers. To upsample an intermediate feature map in turn, e.g., to obtain a result in the original spatial resolution, classical interpolation approaches, such as nearest neighbour or bilinear interpolation, or strided transposed convolutions (Long et al., 2015) are employed. While classical interpolation techniques do not depend on training data, the latter ones introduce weights and biases (comparable to normal convolutions) that need to be trained.

Finally, three concepts from the CNN domain are introduced that are used in this work: encoder-decoder structures (Ronneberger et al., 2015), fully convolutional networks (Long et al., 2015) and residual blocks (He et al., 2016). In an encoder-decoder structure, an input is first downsampled

multiple times to achieve a compressed and highly abstracted representation of this input that considers a large receptive field. Because the objective of this procedure is to obtain one prediction per pixel, the compressed feature map is upsampled in the decoder phase until the original spatial resolution is recovered. To minimise artefacts and smoothing effects caused by computing an output of higher resolution than the input, so-called skip connections are introduced. These connections link feature maps of equal resolution in the encoder and decoder stage and guide the upsampling process. Fully convolutional networks, on the other hand, are characterised by the absence of fully-connected layers. Consisting only of translational invariant components, such as convolutions and pooling, the whole network could be interpreted as a single non-linear filter. This property allows to apply a network on input images of various sizes without the necessity to rescale these images first or to adapt the network architecture to inputs of different sizes, which is typically necessary in the context of conventional MLPs. Lastly, residual blocks consist of one or multiple layers that are bypassed by a skip connection. The output of such a residual block is the sum of the input feature map of this block and the result of the last layer skipped. This simplifies the learning task of the skipped layers to a residual mapping from the features provided as input to those desired as output. As a result, the problem of vanishing gradients can be avoided, enabling the training of much deeper networks.

GC-Net Architecture

In this section, the architecture of the Geometry and Context Network (GC-Net) presented by Kendall et al. (2017b) is reviewed, which is used as basis for the approach developed in the context of this thesis. GC-Net is an end-to-end trainable CNN architecture consisting of 37 layers that predicts a disparity map from a planar rectified stereo pair of three channel colour images. For this purpose, GC-Net performs four major processing steps: feature extraction, cost volume construction, cost volume optimisation and disparity map extraction. While these four steps are described in detail in the following, an overview of the architecture is provided in Figure 2.4.

First, features are extracted from both images of a pair separately, using two branches of 18 2D convolutional layers arranged in residual blocks. While the very first layer consists of a 5×5 filter kernel with stride 2, all subsequent 2D convolutional layers consist of 3×3 filter kernels with stride 1. All layers of this first part of the network have 32 filter channels. In order to force the network to extract similar features from both images, the weights from both branches are shared which is also referred to as Siamese network architecture. In the second step, a cost volume is built by concatenating a feature vector from the left image with a feature vector from the right image for all potential point correspondences, defined by the corresponding horizontal epipolar line and the specified disparity range. By simply concatenating feature vectors from the two images, the subsequent layers are trained to compute the similarity between these two feature vectors instead of explicitly defining a similarity measure in advance. Due to the stride of 2 in the first layer of the network, the resulting 4D volume has a spatial resolution equal to $\frac{1}{2}$ of the input images.

This initial cost volume is further processed using 3D convolutional and transposed convolutional layers with $3 \times 3 \times 3$ filter kernels arranged in an encoder-decoder structure with skip connections.

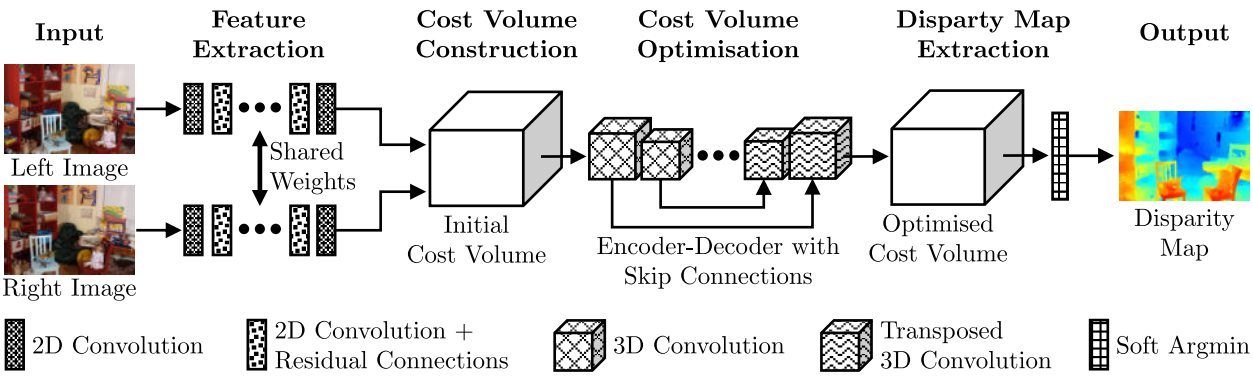


Figure 2.4: **Overview of the GC-Net architecture.** Performing four major processing steps (feature extraction, cost volume construction, cost volume optimisation, disparity map extraction), GC-Net presented by Kendall et al. (2017b) predicts a disparity map from a planar rectified stereo image pair. Source: Adapted from Mehltretter (2020).

In the encoder, the spatial dimension is halved with every third layer using a stride of 2, while the number of feature channels is doubled. Starting with a spatial resolution equal to $\frac{1}{2}$ of the extent of the input images and 32 feature channels, the final encoding is characterised by a spatial resolution equal to $\frac{1}{32}$ of the extent of the input images and 128 feature channels. In the decoder, the spatial resolution is doubled while the number of feature channels is halved with every transposed convolutional layer, until the extent of the resulting feature map is equal to the one of the initial cost volume. The final transposed convolutional layer up-scales this feature map to the spatial resolution of the stereo images pair used as input and reduces the number of filter channels to one. Thus, the output of this last layer is a 3D cost volume similar to the one computed by conventional dense stereo matching approaches (cf. Sec. 2.1). This encoder-decoder structure allows the network to optimise the initial cost volume on different scales and with a large receptive field. In the final processing step, a disparity map is extracted from the optimised cost volume using a differentiable soft argmin layer. All layers apply zero padding and, except for the last 2D as well as the last transposed 3D convolutional layers, are followed by Batch Normalisation (BN) (Ioffe and Szegedy, 2015) and a ReLU non-linearity (Glorot et al., 2011).

The GC-Net architecture demonstrates good accuracy for the tasks of disparity estimation, while having a relatively low number of parameters (∼2.8 Mio.), mainly justified by the absence of fully-connected layers. Moreover, due to its fully-convolutional nature and the final soft argmin layer, GC-Net can process stereo image pairs of varying sizes and can consider different disparity ranges without the need for retraining. Justified by these advantages, the architecture of GC-Net serves as functional model of the BNN developed in the context of this thesis which is presented in Section 4.3.

2.3.2 Bayesian Neural Networks

In contrast to deterministic forms of neural networks, as described in the previous sections, BNNs do not aim to learn point estimates as weights and biases, but probability distributions from which the values of these parameters are randomly sampled anew as part of every forward pass (see

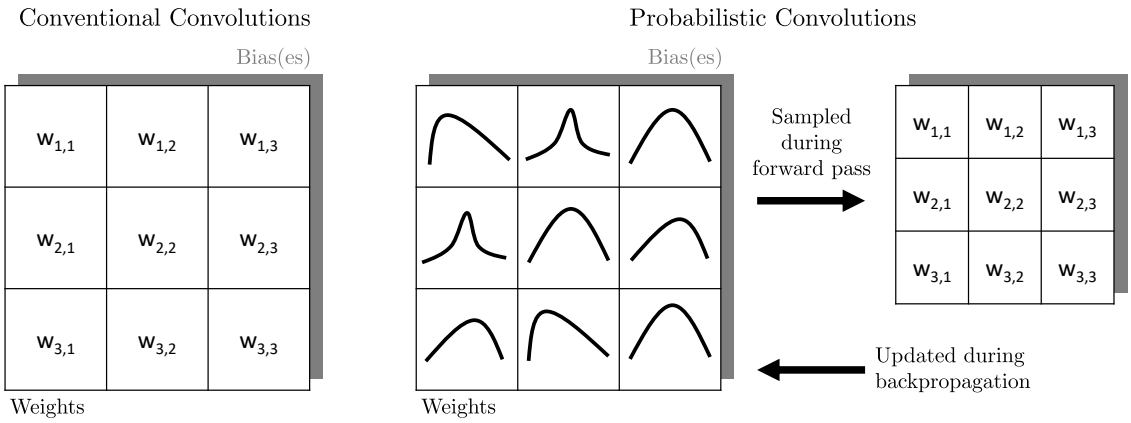


Figure 2.5: **Comparison of conventional and probabilistic convolutional filter kernels.** Conventional convolutions learn point estimates as parameters representing weights and biases. In contrast, probabilistic convolutions learn probability distributions from which weights and biases are sampled during a forward passes.

Fig. 2.5). Thus, from a more formal perspective, a BNN can be understood as a stochastic neural network which is trained using Bayesian inference (Jospin et al., 2020). This procedure is based on the Bayes theorem, aiming to learn the posterior $p(\theta|\mathcal{D})$ of a network's parameters θ under the assumption of a certain prior distribution $p(\theta)$ using data $\mathcal{D}$:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(y|x,\theta)p(\theta)}{\int p(y|x,\theta)p(\theta)d\theta}\,, \tag{2.6}$$

where $p(\mathcal{D}|\theta)$ is the likelihood, $p(\mathcal{D})$ the evidence and $(x, y) \in \mathcal{D}$ are data samples produced by the actual functional dependency $y = \mathcal{F}^*(x)$ which is to be approximated with the defined neural network (see Eq. 2.5). Using this posterior distribution of the network parameters to compute a prediction for a certain data sample, BNNs are able to estimate the epistemic uncertainty associated to this data sample that results from the model learned. Consequently, BNN-based approaches are able to identify samples outside of the training distribution, indicating the limitations of a learned model to process certain data samples. Moreover, the Bayes theorem offers a natural way to implicitly consider and represent commonly applied concepts in the field of deep learning, such as regularisation (Graves, 2011) or ensemble learning (Dietterich, 2002).

Variational Inference

Although BNNs offer a set of advantages and desirable characteristics, computing and thus also sampling from the posterior distribution $p(\theta|\mathcal{D})$ as defined in Equation 2.6 is typically an intractable problem. The reason is the integral involved in the evidence which in general cannot be solved analytically, especially in the context of neural networks which are often characterised by a large number of parameters. To overcome this problem, the exact posterior distribution is commonly approximated circumventing the computation of the evidence, for example, via Markov Chain Monte Carlo (MCMC) approaches (Geyer, 2011) or using VI (Blei et al., 2017). While MCMC approaches allow to sample from the exact posterior distribution directly and are very popular for various applications based on Bayesian statistics, they typically do not scale well with respect

to the model size. Thus, these kind of approaches are not well suited for the domain of neural networks, where the model size is often relatively large (Jospin et al., 2020).

VI, on the other hand, does not sample from the exact posterior distribution, but from an approximation q. To obtain this approximation, also referred to as variational distribution, a family of distributions, such as multivariate Gaussians (Graves, 2011), is chosen over the latent variables of a model, i.e., the network parameters. In order to optimise the variational distribution in the sense that it actually approximates the exact posterior, stochastic variational inference (Hoffman et al., 2013) can be used, which is an adaptation of stochastic gradient descent as commonly employed in the field of deep learning. Together with techniques such as Bayes by Backprop (Blundell et al., 2015) and the reparametrisation trick (Kingma et al., 2015), VI can be easily integrated into an arbitrary deep learning framework, allowing to apply common optimisation algorithms as introduced earlier in this section.

Finally, using VI, the prediction of a BNN is commonly realised via Monte Carlo sampling. For this purpose, n forward passes are carried out for the same input, each of them with a set of parameters $\theta_k \sim q$ with $k \in \{1, .., n\}$ sampled from the approximated posterior distribution q. Combining the results of all these individual forward passes, relevant statistics of the distribution of the prediction, such as the mean or standard deviation, can be computed.

Kullback-Leibler Divergence and Evidence Lower Bound

While VI allows to approximate the exact posterior distribution without the need to compute the intractable evidence, optimising the variational distribution implies to find a set of variational parameters ϕ that minimises the difference between the exact and the approximated distribution. Often, the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) is used to measure the closeness of these two distributions:

$$KL(q_\phi||p) = \int q_\phi(\theta) \log\left(\frac{q_\phi(\theta)}{p(\theta|\mathcal{D})}\right) d\theta = \mathbb{E}\left[\log \frac{q_\phi(\theta)}{p(\theta|\mathcal{D})}\right] . \tag{2.7}$$

However, measuring the difference between two entities commonly requires to know their characteristics, meaning that one would need to know the exact posterior distribution in order to estimate the similarity of the approximation, as it is implied by the presence of the exact posterior $p(\theta|\mathcal{D})$ in the formulation of the KL divergence. While this requirement leads back to the beginning of the problem, instead of minimising the KL divergence directly, it is possible to reformulate this problem using the Evidence Lower Bound (ELBO) instead (Bishop, 2006):

$$ELBO = \mathbb{E}[\log p(\theta, \mathcal{D})] - \mathbb{E}[\log q_\phi(\theta)] = \log p(\mathcal{D}) - KL(q_\phi||p) . \tag{2.8}$$

It can be observed that the ELBO is equal to the negative KL divergence plus a constant $\log p(\mathcal{D})$ that is independent of the variational distribution q_ϕ. Consequently, finding a set of variational parameters ϕ that minimises the KL divergence between the exact posterior distribution $p(\theta|\mathcal{D})$ and an approximation $q_\phi(\theta)$ is equivalent to finding a ϕ that maximises the ELBO (Jospin et al., 2020).

3 Related Work

In this chapter, literature relevant for this thesis is reviewed and discussed. For this purpose, works on dense stereo matching are addressed first in Section 3.1. After this more general view on the topic, a detailed analysis of uncertainty estimation in this area follows. Because of the major differences in their procedure, approaches to estimate aleatoric and epistemic uncertainty are discussed separately in Sections 3.2 and 3.3, respectively. This chapter closes with a discussion of the reviewed literature and a summary of open research questions in Section 3.4.

3.1 Dense Stereo Matching

Although the methodology presented in this work solely focuses on the development of deep learning-based approaches, hand-crafted (sometimes also referred to as expert-based) approaches are also discussed in this section, in order to provide a comprehensive overview of the field. Based on the taxonomy proposed by Scharstein and Szeliski (2002), which is described in Section 2.1, the development from such hand-crafted to deep learning-based approaches is retraced and analysed, before focusing on current CNN-based solutions learning the task of dense stereo matching in an end-to-end manner. This allows to see parallels in both types of approaches and thus helps to understand certain design decisions made in the context of deep learning-based techniques that are based on the knowledge gained from decades of stereo matching-related research.

Cost Computation and Aggregation

According to the taxonomy proposed by Scharstein and Szeliski (2002) and under the assumption that the stereo image pair to be processed is already planar rectified, typically the first stage of dense stereo matching is the computation of matching costs. For this task, various approaches have been presented in the literature that address different application scenarios and have varying strengths and weaknesses. Normalised Cross-Correlation, for example, is robust against Gaussian noise in the images and is suitable to match affine-transformed intensity or colour values, but tends to blur depth discontinuities (Heo et al., 2011). The metric proposed by Birchfield and Tomasi (1998), on the other hand, is insensitive to image sampling (Scharstein and Szeliski, 2002), due to the comparison of a pixel's value in one image with a linear interpolation of the values along the epipolar line in the other image instead of simply comparing intensities pixel by pixel. However, this approach is sensitive to radiometric changes (Heo et al., 2011). In contrast, non-parametric measures, like the Census transformation (Zabih and Woodfill, 1994), do not carry out matching

using the intensities of pixels directly, but determine a value or descriptor for each pixel based on the order of the intensities within a local neighbourhood first. Thus, such measures are typically robust against intensity variations (Hirschmuller and Scharstein, 2009; Stentoumis et al., 2014), but sensitive to noise, especially if the centre pixel of the considered local neighbourhood is affected (Luan et al., 2012). As a consequence of these diverse strengths and weaknesses, several methods were presented in the literature that combine multiple such metrics to compute matching costs in a more robust way (Hu and Mordohai, 2012; Stentoumis et al., 2014). Analysing these works, it becomes clear that a suitable aggregation scheme is essential, in order to retain the strengths and minimise the influence of weaknesses of individual metrics. For this purpose, the uncertainties related to the individual approaches are commonly used in order to fuse the information originating from different cost metrics (Batsos et al., 2018; Mehltretter et al., 2018).

Deep learning-based approaches follow a different idea and learn characteristics important for the decision whether two pixels match or not implicitly from training data, instead of specifying them explicitly by hand. This is commonly achieved using a Siamese feature extractor, a network with two identical branches of convolutional layers, extracting features from the left and the right image, respectively (Chen et al., 2015; Zbontar and LeCun, 2016). By sharing the parameters between these two branches, it is ensured that the features extracted from both images are similar, which facilitates the actual matching. To determine the disparity of a pixel, the corresponding patch is compared against all potentially matching patches from the second image pair-wise, in the sense that every comparison is carried out independently. Luo et al. (2016) argue that the results for a given patch resulting from such a pair-wise matching approach are hard to compare which makes it difficult to determine the correct match. Instead, the authors propose to handle matching as a multi-class classification task, understanding every potentially matching patch together with the associated disparity as one class. To further improve the differentiation whether two patches match or not, Zhang and Wah (2017) propose to learn a specific feature descriptor and require consistency between the descriptors of patches corresponding to the same object point and discriminability between such corresponding to different points (cf. unique matching assumption in Sec. 2.1.3). The downside of all of these approaches is that they are optimised for the identification of correct correspondences on patch-level only, not taking into account a wider range of contextual information. Consequently, as for many hand-crafted approaches, a subsequent global optimisation stage is necessary to achieve accurate results. However, the extracted features may be sub-optimal from the perspective of global optimisation, as only dominant information is retained which is potentially redundant when considering a larger context, while less important but unique information might be discarded in the first stage.

To consider information from more than just one pixel or to aggregate the computed costs and thus to minimise the influence of noise and local outliers, a local neighbourhood is frequently defined as rectangular region around the pixel of interest. While this kind of neighbourhood is easy and fast to compute, it does not take object boundaries into consideration and thus typically smooths over depth discontinuities. More accurate but also computationally more demanding is the approach of cross-based neighbourhoods (Zhang et al., 2009; Zbontar and LeCun, 2016), which decides whether a pixel is considered as being part of a neighbourhood based on the spatial distance to the pixel of interest as well as based on the difference of the colour values assigned. This reduces

the risk of averaging the costs of pixels that belong to different objects and thus may also be located at different depths. Guided filter-based aggregation schemes as, for example, proposed by Hamzah et al. (2017), implement a similar idea, by utilising the intensity gradients of the reference image to determine the shape of the neighbourhood. In contrast, deep learning-related approaches often follow a multi-scale approach to consider multiple neighbourhoods of different extent for every pixel. For this purpose, Chen et al. (2015) extracts two patches of the same size and orientation from the reference image, one in full and one in half resolution. The authors argue that the full-resolution patch facilitates the extraction of sharper depth discontinuities, while the half-resolution patch provides information from a wider context and may thus support matching in weakly textured ares. Park and Lee (2017) and Schuster et al. (2019) follow the same concept and extract patches from multiple scales using a pyramid pooling layer and parallel dilated convolutional layers, respectively. While all these concepts help to mitigate certain effects, such as noise, in a local context, they lack information from global context that is often necessary to resolve ambiguities. Consequently, despite the consideration of local neighbourhood information and the application of local cost aggregation, a subsequent global optimisation is typically necessary to achieve highly accurate and reliable results.

Optimisation

As already discussed before, the consideration of information from a global context and the joint optimisation of all disparity assignments are commonly essential parts of dense stereo matching techniques, crucial to achieve highly accurate results. In this context, global optimisation is commonly interpreted as an energy minimisation problem, defining the energy as a combination of a data term and a smoothness assumption (see Sec. 2.1.2). While the data term considers the matching costs discussed in the previous section, smoothness is achieved by forcing adjacent pixels to satisfy predefined constraints. Such smoothness assumptions may be of first (Hirschmuller, 2008) or second order (Zhang et al., 2014; Kuschk, 2019) preferring fronto-parallel or, more generally, continuous surfaces, may be defined on pixel- or super-pixel-level (Wei and Quan, 2004; Yamaguchi et al., 2014) and may be related to primitives such as planes (Li et al., 2016a), triangles of a mesh (Bulatov et al., 2011) or to more complex objects such as 3D models fitted into the reconstructed scene (Guney and Geiger, 2015). For the optimisation of the resulting energy term, various approaches are presented in the literature: Approaches minimising the energy directly in 2D space, for example, based on Graph-Cuts (Hong and Chen, 2004; Taniai et al., 2014; Li et al., 2016b) or Belief Propagation (Felzenszwalb and Huttenlocher, 2006), result in accurate solutions, but are computationally expensive. On the other hand, one-dimensional approximations, such as Dynamic Programming, are inherently considerably faster, but generally yield results with systematic errors. A common such error case are streaking artefacts in the final disparity map, caused by jointly optimising the disparity estimates of pixels along a single image dimension only, while treating the information along the second dimension as being independent. A well-established and frequently employed trade-off is Semi-global matching (SGM), an approach originally proposed by Hirschmuller (2008). SGM approximates a two-dimensional optimisation by combining multiple 1D scanlines defined in different directions that are applied to the initial cost volume. As a result,

the number of artefacts is significantly reduced, while the computational complexity is kept low. A common problem of such optimisation approaches is the often static and uniform definition of the cost aggregation scheme and the smoothness assumption for all pixels. Properties such as a particularly good matchability due to a unique pattern in an image or, conversely, ambiguous correspondences are typically not explicitly taken into account. Additionally, it is commonly assumed that scenes are piece-wise smooth defining such pieces based on features extracted from the reference image. In this context, relations such as the consistency of image gradients and depth discontinuities (see Sec. 2.1.3) are assumed that may not always hold true, potentially leading to large errors.

Machine learning-based approaches related to the optimisation stage of the taxonomy mainly address these shortcomings: Spyropoulos et al. (2014) as well as Park and Yoon (2015) present random forest-based approaches to learn the identification of reliable disparity estimates from data. Such estimates are used to introduce soft constraints on the neighbourhood relationships into the cost volume or to flatten the cost curves of pixels with unreliable disparity estimates, respectively. Thus, the influence of pixels with reliable disparity estimates on their neighbourhood is implicitly increased during the optimisation procedure. Seki and Pollefeys (2016) propose to identify such reliable disparity estimates using a CNN and to adjust the penalties in the smoothness term of SGM according to the level of reliability assigned. The approaches proposed by Seki and Pollefeys (2017) and Schönberger et al. (2018) also address SGM, but focus on the actual optimisation. They propose to improve the aggregation along individual scanlines and the consolidation of disparity proposals from different scanlines using a CNN or a random forest, respectively. The approach of Jie et al. (2018), on the other hand, introduces a constraint into the optimisation that forces the disparity maps corresponding to the left and the right image of a stereo pair to be consistent. For this purpose, cost volumes referring to both images are computed and refined jointly, using a Recurrent Neural Network (RNN). Carrying out multiple iterations of refinement, the number of unreliable pixels violating the formulated assumption can be reduced, thus increasing the overall quality of the estimated disparity map. While such machine-learning supported optimisation approaches commonly outperform purely hand-crafted ones, the estimation of the trustworthiness of a pixel's disparity estimate is often limited to a small set of properties defined on a subset of all information available in the overall dense stereo matching pipeline. This issue is discussed in more detail in Section 3.2. Moreover, these approaches only address the optimisation stage of the taxonomy leaving the cost computation untouched. Consequently, methods employing such approaches tend to achieve (almost) optimal results for every individual stage, but due to the absence of an overarching optimisation objective, they often only achieve sub-optimal results in total.

Refinement

According to the taxonomy, dense stereo matching is completed by a disparity refinement stage, also referred to as post-processing. Approaches associated with this stage seek to improve the overall accuracy of a previously computed disparity map by detecting and removing or replacing local outliers and by enhancing the remaining disparity assignments. Commonly applied techniques are, for example, the left-right consistency check, which ensures that the disparity maps corresponding to

the left and the right image assign the same depth to pixels showing the same object point, filling holes of pixels without a disparity assignment (Hirschmuller, 2008; Zhang et al., 2017), median and bilateral filtering as well as sub-pixel refinement (Shimizu and Okutomi, 2005; Stentoumis et al., 2014). In the context of machine learning, approaches addressing this stage are commonly defined as residual learning tasks, in the sense that these approaches are trained to predict the difference between the estimated and the ground truth disparity per pixel, for example, using a CNN (Stucker and Schindler, 2020), a RNN (Batsos and Mordohai, 2018) or a variational network (Knöbelreiter and Pock, 2019). For this purpose, besides the initial disparity map, the reference image (Batsos and Mordohai, 2018), both images of a stereo pair (Stucker and Schindler, 2020) or a separately computed uncertainty map (Knöbelreiter and Pock, 2019) may be considered. In contrast, Tosi et al. (2019) implement a different approach, first training a random forest classifier to distinguish between reliable and unreliable disparity estimates. Only keeping reliable estimates from the initial disparity map, unreliable ones are replaced by an interpolation considering reliable disparity estimates located in a local neighbourhood. For this purpose, multiple search lines in different directions are defined and interpolation is applied based on a weighting scheme considering the spatial distance between pixels as well as the difference of their colours in the reference image, similar to the concept of cross-based aggregation described earlier. Finally, Gidaris and Komodakis (2017) propose a combination of these two approaches: Differentiating between inaccurate and incorrect disparity estimates, they learn to refine the first type while replacing the latter one using a CNN. The major challenge for such post-processing approaches is that they are not part of the optimisation and do typically not have access to all information contained in the optimised cost volume. Instead, they are limited to process the estimated disparity map, eventually along with some additional information as discussed before. Consequently, gross errors that were made in previous stages may be impossible to be detected and corrected, for example, because information on ambiguities that occurred during the optimisation is discarded when selecting the most probable disparity estimate to be part of the final disparity map.

End-to-End Learned Approaches

As shown by the literature reviewed so far, the implementation of individual stages of the dense stereo matching pipeline based on deep learning concepts commonly surpasses approaches that are purely based on hand-crafted components, which demonstrates the great potential of this set of techniques. However, optimising the individual stages independently may lead to results that are optimal with respect to the respective stages only, but not for the overall task. Missing an overall optimisation objective, the requirements of approaches addressing later stages are typically not considered in the optimisation of previous ones. To overcome this limitation, several works in the literature propose to interpret the task of dense stereo matching holistically by learning it in an end-to-end manner. For this purpose, all stages of the taxonomy are optimised jointly, commonly using a planar rectified stereo image pair as input to estimate the disparity map referring to the left image directly. In the context of supervised learning, the overall optimisation objective typically addresses the accuracy of this estimated disparity map and demands to minimise its deviations with respect to a reference.

While the concept of end-to-end learning may initially gives the impression that it breaks with the taxonomy of the classical pipeline, the network architectures of most approaches are in fact inspired by this taxonomy. In a first step, a feature vector is generally computed for every pixel of the left and the right image using a Siamese network structure as proposed by Zbontar and LeCun (2016), which was already discussed before. Some works in the literature propose to extend this idea, for example, by extracting features from different scales (Liang et al., 2018) or by enlarging the receptive field using dilated convolutions (Kang et al., 2019) or spatial pyramid pooling (Chang and Chen, 2018). Feature vectors from the left and right image are then combined for every potentially matching pair of pixels, for example, by correlating Mayer et al. (2016) or concatenating Kendall et al. (2017b) them. Combining feature vectors by correlation, the function to compute matching costs is defined explicitly and the result is thus equivalent to the initial 3D cost volume obtained from the conventional cost computation step (cf. Fig. 2.2). Concatenation-based approaches, on the other hand, result in a four-dimensional structure, because potentially matching pairs of pixels are not described by a single cost value, but by their combined feature vector. In this scenario, the function to compute actual matching costs is learned implicitly in the subsequent convolutional layers along with the optimisation of this initial cost volume. For the purpose of optimisation, 2D (Mayer et al., 2016; Pang et al., 2017; Liang et al., 2018) or 3D convolutional layers (Kendall et al., 2017b; Yu et al., 2018) are commonly employed in an encoder-decoder manner, depending on whether the employed feature combination strategy results in a 3D or 4D volume. More sophisticated approaches define explicit aggregation schemes (Cheng et al., 2019), for example, oriented on conventional optimisation techniques, such as SGM (Zhang et al., 2019). Finally, disparities are typically estimated based on the optimised cost volume with sub-pixel accuracy, for example, using directly the output of a convolutional layer (Mayer et al., 2016) or via a differential version of the argmin operation (Kendall et al., 2017b). The ability to examine end-to-end learned approaches using the classical taxonomy not only facilitates the interpretability of these complex neural networks, but also allows for a uniform view on conventional and deep learning based methods. Techniques such as geometry-based regularisation (Knöbelreiter et al., 2017) or guided cost aggregation (Poggi et al., 2019) can therefore be applied directly or can at least be easily transferred to approaches from both groups.

In the literature, the different types of the convolutional layers employed in the optimisation stage are commonly used to distinguish between 2D and 3D architectures (Poggi et al., 2021b). On the one hand, 2D architectures are commonly characterised by a smaller memory footprint as well as less computational effort, due to the lower dimensionality of the employed convolutional layers. On the other hand, 3D architectures enable encoding geometric properties and relationships and provide a higher flexibility, allowing to learn the function for cost computation rather than defining it explicitly. As a consequence, the number of parameters differs significantly between 2D and 3D architectures: For example, the two-dimensional DispNetC proposed by Mayer et al. (2016) has about 36 million parameters, while the three-dimensional GC-Net proposed by Kendall et al. (2017b) has about 2.8 million parameters.

In contrast to classical approaches, end-to-end learned ones often directly include refinement or post-processing procedures. For this purpose, commonly the idea of residual learning is implemented, trying to predict the difference between an initial disparity estimate and the respective

reference value (Pang et al., 2017). Ilg et al. (2018) and Liang et al. (2018) push this idea further and propose iterative refinement schemes that predict and apply residuals multiple times in an incremental procedure. Khamis et al. (2018) and Tonioni et al. (2019) follow a similar approach, but realise an image-pyramid-based idea that estimates disparity on a coarse resolution first and iteratively refines the results by increasing the resolution step-wise, until the original resolution of the reference image is reached. While such an approach reduces the computational effort, it is, like any such step-wise procedure, prone to errors made at an early stage that may not be recoverable afterwards.

For the sake of completeness, it should be mentioned that also several methods exist that learn dense stereo matching along with other tasks, commonly referred to as multi-task learning. Common examples are combinations with semantic segmentation (Yang et al., 2018; Zhan et al., 2019; Wu et al., 2019) and optical flow (Jiang et al., 2019). The results of such multi-task methods demonstrate that it can be beneficial to optimise several tasks jointly taking into account additional aspects such as semantics. However, such approaches are considered to be out of the scope of the present work and are therefore not considered in the subsequent discussions.

While the concept of learning dense stereo matching end-to-end using CNNs leads to convincing results, some limitations of conventional approaches remain open challenges and are complemented by new deficiencies resulting from the end-to-end learning paradigm. Erroneous disparity estimates, for example, still occur frequently in weakly textured areas and close to depth discontinuities. Kang et al. (2019) address this weakness and introduce an auxiliary term to the loss function which aims to minimise the difference between the first deviations of the estimated and the ground truth disparity map. Song et al. (2018), on the other hand, try to predict depth discontinuities based on the intensity gradients of the reference image. While such image information is also available during testing, its consideration to estimate the geometry of the depicted scene is based on the assumption that image gradients and depth discontinuities coincide. However, as already discussed earlier, this is not always the case, therefore such an assumption may be a potential error source itself (cf. Sec. 2.1.3). In addition, deep learning-based approaches are often rather inefficient with respect to their memory consumption and runtime, which is especially true for methods utilising 3D convolutions. To minimise the computational burden, Tulyakov et al. (2018) propose to compress the concatenated feature vectors from the left and right image before forwarding them to the encoder-decoder structure. Arguing that these features initially contain redundant information, the compression reduces the memory footprint without decreasing the accuracy of the obtained disparity map. The approach of Guo et al. (2019) follows a similar idea and clusters the information contained in the feature vectors into groups. The information from the left and the right image is then correlated group-wise, resulting in a vector of correlation scores with a size equal to the number of groups. Because the number of groups is smaller than the size of the initial feature vectors, processing the vector of correlation scores in the encoder-decoder structure requires less computational effort and less memory. While also addressing the efficiency, Duggal et al. (2019) propose a completely different procedure. Based on the concept of the PatchMatch algorithm (Bleyer et al., 2011), a RNN is trained to iteratively select random disparity proposals which are propagated locally. The most promising ones are kept to influence the sampling of proposals in

the subsequent iteration. Because only a subset of all potential disparity proposals is commonly evaluated, the runtime as well as the memory footprint can be reduced significantly.

The last but probably most challenging problem of end-to-end learned approaches arises from the often huge amount of training data required and the strong dependency on this data. As a consequence, such approaches designed for the task of dense stereo matching are commonly pre-trained on large synthetic datasets (Mayer et al., 2016) before fine-tuning them on data from the application domain. Such fine-tuning is typically necessary, because this type of approach is generally sensitive to variations contained in the data to be processed, in the sense that a specific characteristic can only be handled if it has been learned during training. As a result, the performance may decrease significantly if a domain gap exists between training and test data (Poggi et al., 2019; Tonioni et al., 2017). Due to the high range of variations within the real world, however, it is probably impossible to provide divers enough training data to learn all relevant characteristics necessary to reliably operate in a real world scenario (Zendel et al., 2017). While first approaches exist that address this limitation, for example, aiming to learn domain invariant features (Zhang et al., 2020), it remains unsolved and highly challenging.

3.2 Aleatoric Uncertainty Estimation

To estimate the uncertainty inherent in a process, a functional and a stochastic model are commonly needed, independent of the type of uncertainty to be quantified. While the stochastic model defines how uncertainty is understood in a certain context, for example, as a specific probability distribution described by a set of parameters, the functional model specifies how the values of these parameters are determined based on information inherent in the input data. This differentiation is also reflected by the structure of this section, investigating approaches to extract information valuable for the estimation of aleatoric uncertainty first, before different stochastic models to actually quantify aleatoric uncertainty are reviewed.

Functional Model

Initially, the estimation of aleatoric uncertainty was purely based on hand-crafted features defined on the input images or on various (intermediate) representations of the dense stereo matching procedure, such as a cost volume or a disparity map. As shown by Hu and Mordohai (2012) and Poggi et al. (2021a) in comprehensive evaluations, a wide range of such features exist: the properties of a cost curve, such as the curvature or the distinctiveness of the global minimum, the consistency between the disparity maps corresponding to the left and the right image, as well as the distinctiveness of a pixel in its local neighbourhood defined on the reference image. In this context, it is noteworthy that in particular methods which are based on the characteristics of cost curves show convincing results, which is further supported by Veld et al. (2018). This observation can be justified by the higher amount of information contained in a cost volume compared to a disparity map. While a disparity map contains only the supposed optimal disparity for every pixel, a cost volume additionally allows to set this value corresponding to minimal cost into context and

to reason about its actual cost value, its distinctiveness and the ability to localise it. However, each such hand-crafted feature can only consider a rather limited set of characteristics, which restricts its applicability accordingly.

To overcome this limitation and to form more accurate and robust measures, several works propose to combine certain of those hand-crafted features. Beside linear aggregation (Sun et al., 2017), random forest based combinations were especially popular (Batsos et al., 2018; Spyropoulos et al., 2014) considering up to 23 individual features (Haeusler et al., 2013). In this context, the same features are exploited on different scales (Park and Yoon, 2015; Haeusler et al., 2013) and more emphasis is placed on features defined on the disparity image, such as the distance to depth discontinuities (Spyropoulos and Mordohai, 2015; Park and Yoon, 2015) and the disparity distribution within a local neighbourhood (Poggi and Mattoccia, 2016b). Using a random forest allows to learn the importance, and thus the weighting, of individual features from training data. Consequently, such approaches are typically more flexible than predefined linear aggregation schemes and are able to address more divers characteristics than approaches based on a single feature. This idea of learning the combination of several hand-crafted features from training data is pushed further by the works of Seki and Pollefeys (2016) and Poggi et al. (2017), marking the advent of the deep learning paradigm into uncertainty estimation for dense stereo matching. While the authors still use hand-crafted features, neural networks are used to combine these features replacing the previously described random forest component. Despite the large diversity, the employed features are still defined by hand, i.e., they can only evaluate characteristics on an abstract or general level, lacking the ability to consider details or distinguish fine differences.

This issue is resolved by also learning the feature extraction from training data, modelling the whole uncertainty estimation process as a single CNN. For this purpose, Confidence Convolutional Neural Network (CCNN) presented by Poggi and Mattoccia (2016c) as well as the approaches proposed by Poggi and Mattoccia (2016a, 2017) utilise square patches extracted from disparity maps and centred on a pixel of interest to determine the uncertainty associated to the disparity estimate of this pixel. The main idea of this approach is to analyse the disparity distribution of such a patch, expecting smooth surfaces and sharp depth discontinuities rather than high frequency patterns if the estimated disparities are correct. Seki and Pollefeys (2016) extend this idea and propose to use two of such disparity patches, one related to the left, one to the right image of a stereo pair, in order to incorporate the concept of left-right-consistency. In contrast, Fu et al. (2019) suggest an approach called Late Fusion Network (LFN), in which patches from disparity maps and the RGB reference image are both used as input for the network. As a result, additional information on the scene appearance can be considered in the uncertainty estimation procedure. For example, weakly textured areas can be easily identified based on the RGB image, which is relevant since such areas are typically characterised by a higher uncertainty even if the associated disparity estimates appear to be smooth, because of the challenge to localise matching points in such an image region. Moreover, colour or intensity gradients in the reference image may support or contradict the presence of depth discontinuities in the disparity map, allowing to refine the corresponding uncertainty estimates (cf. the relation of image gradients and depth discontinuities in Sec. 2.1.3). With the Local-Global Confidence Network (LGC-Net), Tosi et al. (2018) present an approach that complements the features extracted from a local neighbourhood, as used by the

methods discussed before, with information from the global context. For this purpose, a two-part network architecture is proposed, which uses a local component (Fu et al., 2019; Poggi and Mattoccia, 2016c) to detect details, such as high frequency changes, and an encoder-decoder-based module to enlarge the receptive field. In summary, the CNN-based approaches discussed so far learn to estimate aleatoric uncertainty on features from two-dimensional input only, neglecting the three-dimensional cost volume. However, while processing three-dimensional data poses a computationally more complex task, the additional information provided by such volumes may allow for more accurate estimations, as demonstrated by hand-crafted approaches discussed earlier and confirmed by the findings of Poggi et al. (2021a).

Kim et al. (2017) and Kim et al. (2019b) consider this aspect and propose to learn features on cost volumes. Arguing that in general the cost distributions of raw cost volumes do not allow for direct uncertainty estimation, they propose to preprocess the data. For this purpose, a certain number of candidates having the lowest matching costs assigned are extracted from the cost volume for every pixel, providing only the matching probabilities of these candidates to their network. However, this preprocessing step limits the information provided to the actual uncertainty estimation, potentially preventing the method from exploiting the full potential of learning features on cost volumes. In (Kendall, 2017), on the other hand, uncertainty is learned from the cost volume directly, by passing this volume through a single 3D convolutional layer before taking the average along the disparity axis. While this seems to be a rather simple approach, it has to be mentioned that these two layers are an extension to the GC-Net architecture (see Sec. 2.3.1) enabling the network to learn disparity and uncertainty prediction jointly. Consequently, the cost volume is already optimised to also contain information valuable for the prediction of uncertainty. While Shaked and Wolf (2017), Kim et al. (2019a) and Kim et al. (2020) follow a similar approach to learn disparity and uncertainty estimation jointly, the two latter ones additionally put a focus on learning features from several input modalities, namely the reference image, the disparity map and the cost volume. Such multi-modal input supports learning a higher diversity of features which appears to be beneficial. Despite the fact that all the cost volume-based approaches discussed so far show convincing results, the actual advantage gained by considering cost volumes instead of 2D input is difficult to assess in this context: While the approaches working on information from the disparity map interpret uncertainty estimation as a post-processing step of disparity estimation and thus assume the disparity estimates to be constant, the approaches utilising cost volumes learn the tasks of disparity and uncertainty estimation jointly. Consequently, the input of the uncertainty estimation task, which was assumed to be constant before, can now be adapted during training to be optimally suited for this task. However, this prevents the direct application of such an uncertainty estimation approach on other dense stereo matching methods than the one it was developed for.

Finally, Gul et al. (2019) pursue a completely different approach by using a RNN to learn the task of uncertainty estimation from individual cost curves. In this context, uncertainty estimation is learned independently of the stereo matching task, which means that the cost information is constant during training and that this approach does not suffer from the problem of not being transferable to arbitrary dense stereo matching methods. Using a three-layer architecture including a Long Short-Term Memory block, about 150 trainable parameters are sufficient to achieve results comparable to those of disparity map-based CNN approaches and being even superior with respect

to the ability to generalise to unseen data. This is remarkable, considering that such CNN-based approaches typically have more than a hundred thousand parameters. However, in contrast to the cost volume-based methods discussed before, this approach only processes the cost curve of a single pixel (cf. Fig. 2.2) neglecting adjacent ones. Limiting the information considered in such a way, this approach is vulnerable with respect to noise and outliers in the cost volume.

Stochastic Model

Driven by the fact that, in contrast to depth, typically no ground truth is available for the associated uncertainty, uncertainty estimation has to be learned implicitly. While the difference between the estimated and the ground truth disparity serves as basis for this task, considering only the resulting residuals in the optimisation objective, limits the prediction capability to the estimation of correction terms that may help to refine a initial disparity map. In order to assess the uncertainty in this context, however, a relation between the actual error and the uncertainty has to be established. In the case of aleatoric uncertainty, this typically requires the assumption of a certain stochastic model over the predictions that is valid for all pixels.

Almost all of the methods mentioned in the previous section measure uncertainty in terms of confidence (Hu and Mordohai, 2012; Tosi et al., 2018; Kim et al., 2019b). For this purpose, a score between zero and one is assigned to every pixel, representing the probability of this pixel's disparity estimate being correct. Hence, confidence estimation is often realised as a binary classification task, training a CNN to distinguish between *correct* and *incorrect* disparity estimates, while using the class probability of *correct* as confidence score. The advantage of the confidence-based modulation is its simplicity and the fact that it can be learned as a binary classification task: Specifying an error threshold, the disparity estimates can be distinguished into correct and incorrect based on their deviation from the corresponding ground truth. In this setup, the confidence results as the class probability of a depth estimate being correct. While this approach allows to sort disparity estimates with respect of their probability of being correct and thus, to sort out estimates that are highly likely to be erroneous, it has several downsides: First, a confidence score is always limited in its expressiveness to the specified error threshold, which is often set rather high, e.g., 3 pixels (Poggi and Mattoccia, 2016a). This directly leads to the second problem that confidence does not allow to draw a conclusion about the actual error magnitude other than being potentially higher or lower than this threshold. An assessment of the uncertainty in pixels or metric units is thus not possible. Following from that, the confidence-based model is difficult to integrate into an uncertainty quantification approach combining individual estimates for the aleatoric and epistemic uncertainty.

Alternatively, aleatoric uncertainty can be learned in a Bayesian way via maximum likelihood estimation (Kendall and Gal, 2017). The basic idea of this approach is to understand the parameters of a predetermined type of probability distribution, for example, mean and standard deviation of a Gaussian distribution, as predictions, usually using the estimated disparity as the mean in this context. The predictor, for example, a CNN as discussed in the previous section, is trained to maximise a likelihood function, such that under the assumed distribution the ground truth

disparity, which is used as observation, becomes most likely. While this approach is applied to a diverse field of applications, such as semantic segmentation and monocular depth estimation (Kendall and Gal, 2017), optical flow (Gast and Roth, 2018) and multi-view stereo based on a single moving camera (Vogiatzis and Hernández, 2011), it is far less popular in the context of dense stereo matching compared to the concept of confidence. This might be explained by the requirement of more detailed knowledge on the real error distribution. However, contrary to the concept of confidence, this approach allows to additionally quantify the uncertainty in pixels or metric units.

Kendall (2017) adopts this Bayesian approach for uncertainty estimation in the context of dense stereo matching, assuming a Gaussian distribution over the predictions of his model. By formulating the loss function as the negative log likelihood of the assumed Gaussian distribution, the likelihood is maximised using common deep learning optimisation strategies (see Sec. 2.3). Kendall and Gal (2017) employ the same procedure for the task of monocular depth estimation, but assume a Laplacian instead of a Gaussian distribution. They argue that the L1 norm, which corresponds to the Laplacian distribution, commonly outperforms the L2 norm for vision-related regression tasks. Moreover, it becomes apparent that this approach not only allows to quantify aleatoric uncertainty, but also improves the accuracy of the disparity estimates by acting as a regulariser which weights the disparity error of a training sample by the inverse of the estimated uncertainty. Both, the Gaussian as well as the Laplacian model, assume a uni-modal error distribution expecting a distinct global optimum that is good to localise. However, this is a strong simplification that is often violated in real-world scenarios, e.g., in the context of dense stereo matching due to weakly textured or occluded regions in an image. In contrast, Vogiatzis and Hernández (2011) and Pizzoli et al. (2014) assume a mixture distribution to model the uncertainty inherent in the results of multi-view stereo reconstructed from images captured with a single moving camera. The basic idea of this assumption is that a depth sensor produces two types of measurements: good measurements that are normally distributed around the correct depth and outlier measurements that are uniformly distributed in an interval that contains the correct depth. Both distributions are combined by a weighting factor set individually for each pixel, whereby the parameters characterising this mixed distribution are iteratively adjusted by convex optimisation. Although the assumption of such a mixture distribution better approximates the real error distribution, it has neither been investigated in the context of dense stereo matching nor in a deep learning framework so far. Consequently, the verification whether this approach is suitable to be combined with a deep learning-based functional model, as described in the previous section, is still to be done.

3.3 Epistemic Uncertainty Estimation

Similar to the previous section on aleatoric uncertainty estimation, a differentiation into functional and stochastic model is also possible addressing epistemic uncertainty. However, because epistemic uncertainty describes the uncertainty inherent in a certain method and its parameters, the functional model is defined by the choice of this method, for example, as a specific CNN architecture designed to predict a disparity map from a stereo image pair as discussed in Section 3.1.

Consequently, this section focuses on the possibilities to define the stochastic model, treating the functional model as given. Compared to aleatoric uncertainty, which is commonly treated as an additional predictive value, the estimation of epistemic uncertainty is typically more difficult. However, this type of uncertainty helps to mitigate the problem of overconfident predictions (Gal and Ghahramani, 2016), which occurs in particular in the context of neural networks, and to identify cases in which a method is highly uncertain regarding its prediction, for example, processing data outside of the learned data distribution. To cope with this task, in particular the use of stochastic neural networks has proven to be well suited. In contrast to the commonly employed deterministic neural networks that learn point estimates as parameter values, stochastic ones allow to learn a distribution over the parameters (Jospin et al., 2020). Epistemic uncertainty is then commonly estimated via Monte Carlo sampling. Aggregating the predictions of the individual samples, this procedure allows to approximate the central moments of the probability distribution describing the final result, such as the mean and the variance.

In this context, ensemble learning can be understood as an approximation of such stochastic neural networks. Training a set of structurally identical deterministic networks independently, the predictions of the individual networks are combined at test time to estimate the epistemic uncertainty. The most popular way of setting up such an ensemble is to use different initial values for the network parameters, which is typically achieved by random initialisation with varying seed values (Lakshminarayanan et al., 2017). However, also other types of ensembles exist, such as training individual networks on different subsets of the training data (Breiman, 1996; Moukari et al., 2019) or using the parameter values of the same network obtained after various numbers of training epochs as individual networks to form an ensemble (Huang et al., 2017). Generally, it is argued in the literature that procedures based on ensemble learning require less computational effort than other variants of stochastic neural networks, while still leading to good results (Lakshminarayanan et al., 2017; Ovadia et al., 2019). However, while such an approach may be reasonable for some scenarios, it is not feasible for large architectures, especially if a larger set of networks is to be considered in the ensemble. Because training (and potential fine-tuning) is commonly carried out independently, the computational effort grows linearly with the size of the ensemble, also requiring that the parameter values of all networks are present at test time, leading to an enlarged memory footprint. Moreover, ensemble learning does typically not allow to consider prior knowledge on the uncertainty or to model correlations between the network parameters or the individual instances forming an ensemble.

A second realisation of the concept of stochastic neural networks is Monte Carlo dropout, which is, for example, used by Gal and Ghahramani (2016), Kendall et al. (2017a) and Kendall and Gal (2017). Similar to the variant of dropout commonly used for the purpose of regularisation during training (Srivastava et al., 2014), Monte Carlo dropout places a Bernoulli distribution over the network weights setting them to zero with a certain probability. However, the Monte Carlo variant applies this procedure not only during training but also at test time. Thus, with every forward pass a slightly different parametrisation of the same network is used to obtain a prediction, leading to varying results. In contrast to most ensemble learning techniques, only the parameters of a single network need to be learned and be present at test time, which clearly reduces the computational effort during training as well as the memory footprint when testing. Furthermore, using Monte

Carlo dropout, the number of forward passes and thus the size of the ensemble can be changed flexibly at test time without the need to train further variants of the network. A potential weakness of Monte Carlo dropout is that the uncertainty predictions are typically not calibrated, in the sense that systematic deviations may exist between the estimated and the actual observable variance in the test data. However, this problem can be mitigated when also learning the dropout probability from the training data, for example, using variational dropout proposed by Kingma et al. (2015). Another limitation is the frequent absence of the possibility to take into account prior knowledge or the correlation between parameters of the network. Especially the latter appears to be problematic in the context of CNNs, since the convolutional filter kernels are spatially correlated. While Ghiasi et al. (2018) have investigated this issue for conventional dropout and presented a solution to improve the regularisation capability by taking into account spatial correlations, the implications of such correlations on the estimation of epistemic uncertainty remain an open research question.

BNNs constitute the third and last realisation of stochastic neural networks being discussed in this section that allows to define a prior for the parameters of the network, treating the uncertainty in a Bayesian manner (cf. Sec. 2.3.2). Despite the fact that the basic concepts of BNNs are already known for decades (MacKay, 1992; Neal, 1995), they have only recently been used in practice for more complex tasks, such as image-based object classification (Brosse et al., 2020). For a rather long time, the complexity of larger BNNs was a major challenge causing training via variational inference (Graves, 2011) to converge slowly and to sub-optimal solutions while requiring high computational effort. Consequently, such approaches were mainly theoretically motivated but had limited relevance in practice. However, more recent advances in the field of variational inference, such as stochastic variational inference (Hoffman et al., 2013), Bayes by backprop (Blundell et al., 2015), the reparametrisation trick (Kingma et al., 2015) and flipout (Wen et al., 2018), have mitigated this problems significantly. With a runtime required for training that is only slightly increased compared to a deterministic baseline, nowadays BNNs can be trained faster than an ensemble of networks. Moreover, this type of stochastic neural networks offers the flexibility to model the distribution over the parameters in various ways, considering prior information, and to also take into account correlations between parameters. In this context, the network parameters are not learned directly, but drawn from a learned variational distribution. While the randomness of the sampled network parameters offers a natural regularisation during training and minimises the risk of over-fitting, learning a variational distribution may lead to a significant increase of trainable parameters. Assuming, for example, that every network parameter is drawn from a Gaussian distribution, two variational parameters would need to be learned per network parameter to parameterise this Gaussian. Thus, already a rather simple type of distribution, such as a Gaussian, doubles the number of trainable parameters compared to a deterministic baseline if the covariance matrix is assumed to be diagonal, i.e., correlations are not considered. Addressing this issue, Zeng et al. (2018) and Brosse et al. (2020) have recently proposed to treat only some layers of the network in a probabilistic manner while keeping all others deterministic. Following this procedure, the number of parameters and the computational effort can be reduced, while achieving comparable results with respect to the prediction accuracy and the quality of the estimated epistemic uncertainty. However, so far this approach is only examined with respect to classification tasks and requires further investigations, especially with respect to regressions tasks such as dense stereo matching.

3.4 Discussion

Based on the findings from the literature reviewed in the previous sections, current limitations are identified and research questions that remain open in the context of dense stereo matching and the estimation of the associated uncertainty are summarised and discussed in this section. The insights of this discussion serve as further motivation of the methodology proposed in this work and are addressed in the subsequent chapter.

Dense Stereo Matching

As revealed by the literature review, the introduction of deep learning marked a milestone in the development of dense stereo matching techniques, increasing the accuracy achieved significantly. This is especially true for approaches that interpret the taxonomy of Scharstein and Szeliski (2002) holistically and implement all of its stages as part of a single consistent neural network that is trained end-to-end. However, recent publications also demonstrate that a number of long-known challenges remain, for example, posed by weakly textured areas, occlusions and repetitive patterns. Moreover, these challenges are accompanied by new ones caused by the mainly data-driven nature of deep learning-based procedures. In this context, the need for a huge set of training data with reference disparity has to be mentioned, however, the sensitivity regarding domain gaps between training and test data poses a vastly larger problem. In case that the training and test dataset belong to different domains, such as indoor and outdoor, commonly a clear decrease of accuracy has to be expected. Consequently, despite the often convincing results of recent dense stereo matching methods, the estimation of the uncertainty associated to a disparity estimate remains an important aspect and is crucial to identify circumstances in which such methods lead to inaccurate or even completely incorrect results. Furthermore, with respect to the previously addressed challenges, the consideration of both aleatoric as well as epistemic uncertainty appears to be crucial. Deep learning-based procedures more strongly require to estimate epistemic uncertainty due to the additional uncertainty of the learned model parameters alongside the model form uncertainty that was already inherent in conventional hand-crafted methods. Bayesian deep learning, in particular, rather opens up new possibilities allowing to accurately model and assess these kinds of uncertainties which was previously not possible or only hardly feasible.

Aleatoric Uncertainty Estimation

Analysing the advantages of the different approaches for estimating aleatoric uncertainty, it is noticeable that learned features outperform hand-crafted ones and that features defined on cost curves demonstrate superior performance. However, most approaches in the literature preprocess cost volumes before providing them to a neural network in order to estimate the uncertainty (Kim et al., 2017; Kim et al., 2019b). While it is argued that such preprocessing steps are beneficial to reduce the computational effort or to increase the robustness of these procedures, they often reduce the amount of information contained in the cost volumes. Thus, information that may be

beneficial for the task of aleatoric uncertainty estimation is neglected, artificially limiting the quality of the resulting uncertainty. Moreover, it is evident that the consideration of multiple cost curves belonging to pixels located within a local neighbourhood leads to a more robust estimation than relying on a single cost curve, as the impact of noise is mitigated in a better way. Consequently, it appears to be reasonable to extract features directly from raw cost volumes taking into account the information corresponding to multiple pixels.

With respect to the stochastic model, most of the works published in the literature focus on the estimation of confidence. While confidence is rather simple to learn by formulating it as a binary classification task, its expressiveness is limited: The predicted probability of a pixel's disparity estimate being correct is always to be understood relative to an error threshold that needs to be specified before training. The actual error magnitude is thus not assessable, which makes it impossible to express the uncertainty in pixels or metric units. Approaches from the field of Bayesian deep learning allow to overcome this limitation by learning the parameters of a probability distribution describing the result (Kendall and Gal, 2017), i.e., in the context of dense stereo matching, a disparity estimate. Thus, aleatoric uncertainty, for example, in form of the variance, is either predicted by a network directly as additional output or can be derived from the predicted disparity distribution. Following this procedure, only uni-modal distributions, such as a Laplacian or Gaussian distribution, are employed in the context of depth prediction based on mono or stereo images so far (Kendall and Gal, 2017; Kendall, 2017). However, such distributions imply the presence of a single and distinct global optimum, an assumption that is not valid for all pixels in the context of matching and is violated, for example, in the case of occluded pixels or pixels located in weakly textured areas of an image. Consequently, the usage of mixture distributions might be beneficial to approximate the real error distribution more accurately.

Epistemic Uncertainty Estimation

With the rise of deep learning and stochastic neural networks in particular, the estimation of epistemic uncertainty became practically feasible. However, most works published in the literature focus on rather simple applications, such as regression problems predicting chemical (Hernández-Lobato and Adams, 2015), physical (Gal and Ghahramani, 2016) or simulated functional dependencies (Blundell et al., 2015) that involve a small number of parameters trained on just a few hundred data samples. In the context of photogrammetry and computer vision, mainly classification problems are addressed, for example, classifying hand-written digits from the MNIST dataset (Kingma et al., 2015; Gal and Ghahramani, 2016; Zeng et al., 2018) or objects from the CIFAR or ImageNet datasets (Lakshminarayanan et al., 2017; Brosse et al., 2020). Lately, also more challenging tasks are investigated, such as monocular depth prediction (Kendall and Gal, 2017; Moukari et al., 2019) or semantic segmentation (Kendall and Gal, 2017; Kendall et al., 2017a). To the best of the author's knowledge, dense stereo matching was not yet subject of investigations in the context of epistemic uncertainty estimation. To cope with these tasks, most works rely either on ensemble learning or on Monte Carlo dropout. However, ensemble learning requires training every network of an ensemble independently and retaining all learned parameters at test time. Thus, at present such a procedure is not reasonable for more complex tasks that require large neural networks, such

as CNNs designed for the purpose of dense stereo matching. Monte Carlo dropout, on the other hand, has demonstrated to work well on a broad variety of applications including complex tasks, without leading to such a computational overhead. However, Monte Carlo dropout is limited to a Bernoulli distribution over the network parameters with tuning the dropout rate being the only option for adaptation. With the recent developments of variational inference, also BNNs achieved practical relevance, yet offering more flexibility to model stochastic properties compared to the other approaches discussed so far. With the ability to consider prior knowledge and correlations between different network parameters, BNNs appear to be a powerful tool to estimate epistemic uncertainty that is worth being investigated in more detail from the practical perspective of dense stereo matching.

4 Uncertainty Estimation for Dense Stereo Matching - A New Method

In this chapter a novel method to estimate aleatoric and epistemic uncertainty in the context of dense stereo matching is proposed. Starting with an overview including a description of the input data and the problem statement in Section 4.1, approaches to estimate aleatoric (Sec. 4.2) and epistemic uncertainty (Sec. 4.3) are first introduced separately before proposing a way to combine them in Section 4.4. This chapter closes with a discussion in Section 4.5, addressing assumptions made in the context of this approach as well as limitations of the presented methodology.

4.1 Overview

The overall objective of the method proposed in this work is the estimation of uncertainty associated to disparity estimates obtained via dense stereo matching. For this purpose, both aleatoric as well as epistemic uncertainty is considered, which arise from uncertainty contained in the data or is inherent in the specified model and its parameters, respectively. The input to the proposed method are stereoscopic image pairs $(\mathbf{I}_L, \mathbf{I}_R)$, referring to the left image $\mathbf{I}_L$ of such a pair as the reference image. It is assumed that both images were captured simultaneously, allowing to neglect the influence of movements of parts of the scene depicted, and have a reasonable overlap in which the depth can be determined via triangulation. Moreover, it is assumed that the stereo setup is calibrated, which implies that the interior orientations of both cameras and the relative orientation between them is known. Lastly, the stereo image pairs are presented to the proposed method after planar rectification, resulting in epipolar lines that coincide with the image rows (cf. Sec. 2.1.1).

The expected output is a tuple, consisting of a disparity map $\mathbf{D}$ and an uncertainty map $\mathbf{U}$, which correspond to the reference image, having the same size, and containing a disparity estimate $d_\mathbf{p}$ and a variance $\sigma^2_\mathbf{p}$ for every pixel $\mathbf{p}$ of the reference image, respectively. Thus, the functional relationship between input and output is described as:

$$f(\mathbf{I}_L, \mathbf{I}_R) = (\mathbf{D}, \mathbf{U})\,. \tag{4.1}$$

To allow for separate investigations, this function is further subdivided into three main processing steps: cost volume construction $f^{(c)}$, disparity estimation $f^{(d)}$ and aleatoric uncertainty estimation $f^{(a)}$. In this context, $f^{(c)}$ includes all stages of the classical taxonomy of Scharstein and Szeliski (2002) that relate to the cost volume, in particular, the cost computation and cost aggregation stages as well as a global optimisation, potentially carried out on the cost volume. The extraction of the disparity map from the (optimised) cost volume is in turn represented by $f^{(d)}$. Note that

no post-processing techniques are investigated or applied in the context of this work, which is the reason for neglecting this stage in the functional relationship presented. Lastly, the uncertainty map $\mathbf{U}$ is assumed to be an aggregation of the aleatoric and epistemic uncertainty maps $\mathbf{U}_A$ and $\mathbf{U}_E$ corresponding to the disparity estimation procedure described. These considerations lead to the following refined functional relationship:

$$f^{(c)}(\mathbf{I}_L, \mathbf{I}_R) = \mathbf{C}\,, \quad f^{(a)}(\mathbf{C}) = \mathbf{U}_A\,, \quad f^{(d)}(\mathbf{C}) = (\mathbf{D}, \mathbf{U}_E)\,, \tag{4.2}$$

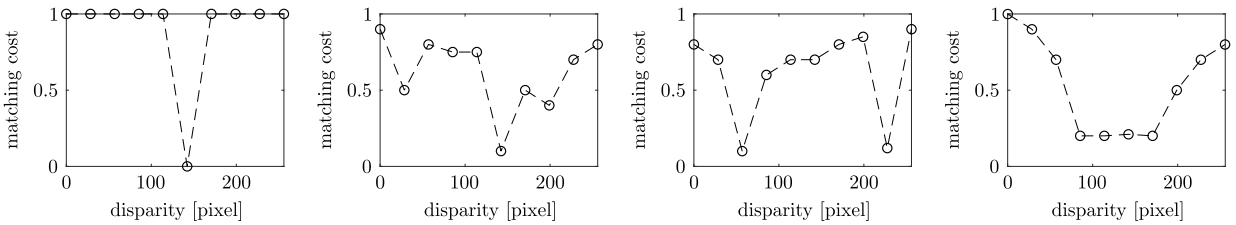
where $\mathbf{C}$ describes a three-dimensional cost volume, corresponding to the reference image, as introduced in Section 2.1.2. In the following sections, the individual steps are first addressed separately, before being fused and considered as a whole in Section 4.4. In more detail, a CNN-based approach that implements $f^{(a)}$ and that is trained with techniques from the field of Bayesian deep learning is presented in Section 4.2, assuming $f^{(c)}$ to be arbitrarily defined. Thus, the proposed approach to estimate aleatoric uncertainty is able to operate on various methods computing matching costs. Contrary, $f^{(c)}$ together with $f^{(d)}$ are realised as a BNN described in Section 4.3, to estimate disparity and epistemic uncertainty jointly. Note that parts of the methodology presented in this chapter have already been published in: (Mehltretter, 2020), (Mehltretter and Heipke, 2021) and (Zhong and Mehltretter, 2021).

4.2 Aleatoric Uncertainty Estimation

The estimation of uncertainty commonly requires a functional and a stochastic model. While the latter defines how uncertainty is understood in a certain context, for example, as a specific probability distribution described by a set of parameters, the functional model specifies how the values of these parameters are determined based on information inherent in the input data. Motivated by the two promising directions of deep learning- and cost volume-based aleatoric uncertainty estimation discussed in the literature review, in this work, a novel CNN architecture is proposed as functional model in Section 4.2.1, allowing to estimate aleatoric uncertainty from 3D cost volumes directly. In addition, two novel approaches defining the stochastic model based on mixture distributions and in a Bayesian way are presented in Section 4.2.2, allowing to train the proposed network with techniques from the field of Bayesian deep learning, i.e., maximising the likelihood of the predicted uncertainty over the predefined probability distribution.

4.2.1 CNN-based Cost Volume Analysis

As stated in the overview, this section focuses on the aleatoric uncertainty estimation step $f^{(a)}$ (cf. Eq. 4.2), assuming the cost computation and disparity estimation steps as arbitrarily defined. Thus, a cost volume to be processed may originate from any cost computation method, for example, from conventional hand-crafted ones such as Census-based block matching (Zabih and Woodfill, 1994) or Semi-global matching (Hirschmuller, 2008) or from deep-learning-based ones such as MC-CNN (Zbontar and LeCun, 2016) or GC-Net (Kendall et al., 2017b). Thus, the approach to estimate aleatoric uncertainty from cost volumes presented in the following, is not limited to a specific method



(a) Ideal curve (b) Distinct global minimum (c) Ambiguous minima (d) Flat minimum

Figure 4.1: **Cost curve-based aleatoric uncertainty estimation.** (a) An ideal curve, characterised by a single minimum with zero cost and all other values being one. (b) A more realistic curve with multiple minima, but a reliably identifiable global minimum. (c) No distinct global minimum is identifiable, making the determination of the correct correspondence unreliable - a typical behaviour in areas with repetitive patterns. (d) The occurrence of a wide and flat minimum is a typical behaviour in non-textured areas and leads to an inaccurate localisation of the correct correspondence. While (a) and (b) are typically assigned a low uncertainty, the uncertainty for (c) and (d) is significantly higher. Source: Mehltretter and Heipke (2019).

computing the matching cost. However, it is assumed that the matching cost within such a volume is normalised to the interval $[-1, 1]$. Besides this normalisation, no further preprocessing of the cost volume is assumed or carried out, minimising the risk to limit the amount of information provided to the proposed CNN. As introduced in Section 2.1.2, the width w and height h of a cost volume correspond to the width and height of the reference image, while the matching cost information is encoded along the z axis via so-called cost curves (cf. Fig. 2.2). Regardless of the cost computation method they originate from, typical characteristics can be observed for such cost curves, which can be used to predict the aleatoric uncertainty associated to a disparity estimate: A disparity estimate with low uncertainty is characterised by a clearly identifiable and unambiguous global minimum (Fig. 4.1a and 4.1b). In contrast, a clearly higher uncertainty is usually present if either no distinct global minimum can be identified (Fig. 4.1c) or if the global minimum is wide and flat, making the localisation of the correct correspondence inaccurate (Fig. 4.1d). Such ambiguous situations commonly occur when trying to match pixels that are occluded in one of the images or that are located in areas with weak or highly repetitive texture. While cost curves help to identify such ambiguous matches, they typically do not provide enough information to detect local outliers, for example, caused by noise or artefacts in the images. To overcome this limitation, it is reasonable to not process individual cost curves, but analyse cost volume extracts, which consist of multiple cost curves corresponding to pixels located within the local neighbourhood around a pixel of interest (cf. Fig. 2.2). Such a procedure combines the advantages of using patch-based and cost curve-based features for the task of aleatoric uncertainty prediction.

To realise this idea and to learn the relationship between aleatoric uncertainty and features defined on the 3D cost volume, a novel CNN architecture referred to as Cost Volume Analysis Network (CVA-Net) is presented in this thesis. This architecture follows the idea of a feedforward network and consists of three main components: neighbourhood fusion, depth processing and uncertainty estimation (cf. Fig. 4.2). A detailed layer-by-layer definition can be found in Table 4.1. As input, the network takes cost volume extracts of size $\hat{w} \times \hat{h} \times |\delta|$. As a result of preliminary experiments with extracts of different sizes, the size of the perceptive field is set to $\hat{w} = \hat{h} = 13$ pixels, providing

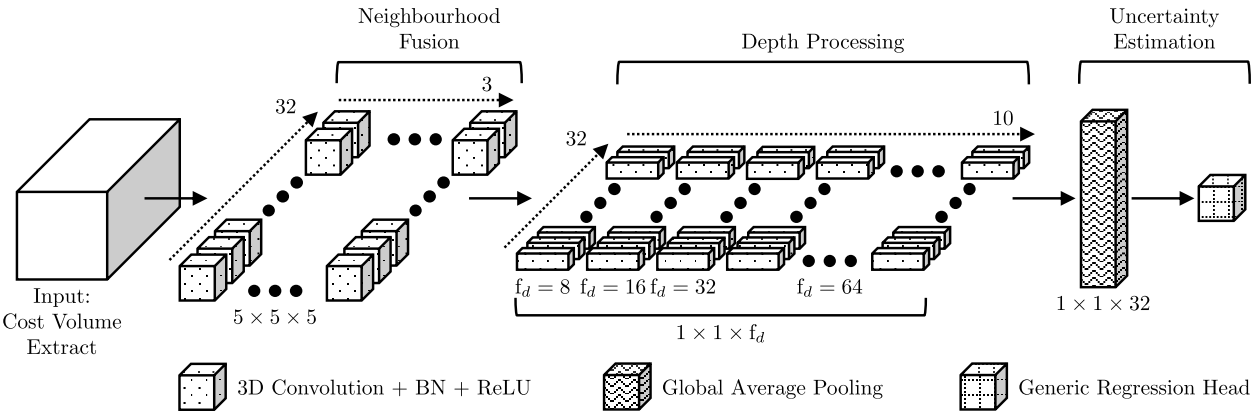


Figure 4.2: **Overview of the architecture of CVA-Net** (Cost Volume Analysis Network). Consisting of three main elements, the network first fuses a cost volume extract into a single cost curve, then this curve is further processed along the disparity axis. These two parts are realised using 3D convolutions, each followed by batch normalisation (BN) and a ReLU non-linearity. While the convolutional kernels have a fixed size in all layers of the first part, the depth of the filter kernels f_d varies in the second part. The average pooling layer together with the regression head at the end of the network allow to estimate the aleatoric uncertainty pixel by pixel independent of the depth of the cost volume extract. The definition of the regression head depends on the stochastic model used, with concrete implementations shown in Figure 4.3. Source: Adapted from Mehltretter and Heipke (2019).

a good trade-off between the amount of information available to the network and the computational effort caused by the employed 3D convolutional layers. In order to process complete cost curves, the depth $|\delta|$ of an extract is chosen to be equal to the depth of the cost volume. Due to the global average pooling layer in the uncertainty estimation stage, CVA-Net is capable of processing cost volumes of variable depth, meaning that no retraining is needed if the disparity range considered by the cost computation method changes. Contrary to the consideration of a fixed cost volume depth, for example, implied by the usage of a fully-connected instead of an average pooling layer, such a varying depth may be beneficial in both directions: Enlarging the depth and thus the search range for potential correspondences in the stereo image pair is reasonable if the depicted scene contains objects that are located very close to the image planes. Such objects have a large disparity, which cannot be determined if the search range considered is too small. On the other hand, reducing the depth of a cost volume and thus the search range lowers the risk of facing ambiguous matching results which typically leads to disparity maps of higher accuracy. Because different scenes are typically characterised by different depth distributions, it is reasonable to be able to adjust the search range dynamically, assuming that a certain amount of prior knowledge on this distribution is available.

The first part of the network, the neighbourhood fusion, merges the information contained in a cost volume extract into a single 1D feature vector using 3D convolutional layers with a filter kernel size of $5 \times 5 \times 5$. The basic idea behind this step is equivalent to that of most region-based matching approaches: Including neighbourhood information increases the robustness. Especially if the cost curve corresponding to the pixel of interest is affected by noise or delivers an ambiguous solution, neighbourhood information may be beneficial. The depth of the 3D convolutional filters

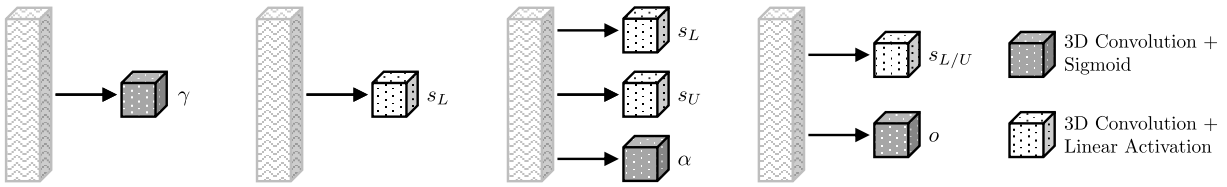
Table 4.1: **Summary of the proposed CVA-Net architecture.** Unless otherwise specified, each layer is followed by batch normalisation (BN) and a ReLU non-linearity. Note that the last layer, the regression head, and the number of its outputs depend on the stochastic model used. See Figure 4.3 for details on the concrete implementations.

Layer	Description	Output Map Dimensions
Input	Cost Volume Extract	$13\times13\times\lvert\delta\rvert$
	Neighbourhood Fusion	
1	3D conv., 5×5×5, 32 filters	$9\times9\times\lvert\delta\rvert-4$
2	3D conv., 5×5×5, 32 filters	$5\times5\times\lvert\delta\rvert-8$
3	3D conv., 5×5×5, 32 filters	$1\times1\times\lvert\delta\rvert-12$
	Depth Processing	
4	3D conv., 1×1×8, 32 filters, zero padding	$1\times1\times\lvert\delta\rvert-12$
5	3D conv., 1×1×16, 32 filters, zero padding	$1\times1\times\lvert\delta\rvert-12$
6	3D conv., 1×1×32, 32 filters, zero padding	$1\times1\times\lvert\delta\rvert-12$
7-13	3D conv., 1×1×64, 32 filters, zero padding	$1\times1\times\lvert\delta\rvert-12$
	Uncertainty Estimation	
14	Global average pooling, linear activation, no BN	1×1×32
15	Generic regression head	various

associated to this part of the network is set to five to handle minor shifts of the curves relative to each other, e.g., caused by discretisation errors which may occur during the cost computation step. The number of layers employed within this part can directly be derived from the size of the perceptive field, because in the specified configuration, every convolutional layer reduces the extents of the feature map by four in every direction.

In the subsequent depth processing part, the resulting 1D feature vector is further processed in order to derive high-level features characterising the cost volume extract. In a simplified way, this part can be interpreted as the extraction of features, such as position and number of local optima and the curvature of the cost curve, which are subsequently used to determine an uncertainty value, as exemplarily shown in Figure 4.1. It is noteworthy that the filter depth f_d increases with the layer depth: Starting with $\mathrm{f}_d = 8$ the value is doubled with every new layer until $\mathrm{f}_d = 64$ is reached. The design with increasing filter depth as well as the number of layers within this part of the network were optimised empirically. However, this design is inspired by the idea that features at a lower level rather consider a local neighbourhood, for example, to identify a local optimum, while features at a higher level rather gather information from a wider context, for example, to aggregate the number and distribution of optima. Furthermore, zero padding is utilised for all convolutions in the depth processing part of the network. This keeps the size of the output feature map constant and, compared to no padding, provides a larger number of features as input for the subsequent uncertainty estimation.

The third and last part of the network consists of a global average pooling layer and a regression head to predict an uncertainty value based on the previously extracted features. The idea behind this part is illustrated in Figure 4.1 in a simplified form. Note that the actual implementation of the regression head depends on the stochastic model used, leading to different definitions of this final layer as well as varying numbers and types of outputs. Details on this layer are thus given



(a) Confidence (b) Uni-modal (c) Mixture (d) Scene-aware

Figure 4.3: **Regression head variants of CVA-Net**. While the basic architecture of CVA-Net is equal for all stochastic models proposed, the regression head, i.e., the final layer, differs. Note that all 3D convolutions in this layer have a kernel size of $1 \times 1 \times 1$. (a) For confidence estimation via binary classification, the confidence γ with $\gamma \in [0, 1]$ is predicted. (b) In the uni-modal case, the log standard deviation s_L of a Laplace distribution is predicted. (c) For the mixture model, log standard deviations s_L and s_U of a Laplace and a uniform distribution as well as the mixture coefficient α with $\alpha \in [0, 1]$ are predicted. Finally, the log standard deviation $s_{L/U}$ either describing a Laplace or a uniform distribution and a region indicator o with $o \in \{0, 1\}$ specifying which type of distribution is assumed are predicted for the scene-aware model.

as part of the descriptions of the different stochastic models in Section 4.2.2, with an overview of all variants shown in Figure 4.3. Due to the absence of fully-connected layers, the proposed CVA-Net architecture is characterised as being fully convolutional. This allows training on image patches while computing an uncertainty map of the full resolution image in a single forward pass at test time. However, this of course also allows for piece-wise processing of the cost volume if hardware restrictions have to be taken into account. With less than 800.000 trainable parameters, the proposed CVA-Net is relatively compact compared to state-of-the-art CNNs designed for the task of aleatoric uncertainty estimation, which often have several million parameters.

4.2.2 Uncertainty Models

The task of disparity estimation from stereo images is commonly learned in a supervised manner, minimising the difference between estimates and corresponding reference data. In the context of uncertainty estimation, such reference data is typically not available, preventing the application of a comparable direct learning procedure. However, under the assumption of a specific uncertainty model, it is possible to learn uncertainty from the distribution of the disparity error, i.e., the deviations between estimated and ground truth disparity, implicitly. While confidence estimation stated as binary classification task is especially popular in the literature, stochastic models based on Bayesian theory have demonstrated to be superior in the sense that they allow the quantification of uncertainty in pixels or metric units (cf. Sec. 3.2). Thus, besides confidence estimation, different variants to set up the stochastic model for the purpose of aleatoric uncertainty estimation in a Bayesian manner are presented and discussed in the following, namely a uni-model variant based on a Laplacian distribution as well as two mixture models, taking outlier measurements or the characteristics of the depicted scene explicitly into account. In addition to different loss functions for each variant, also varying definitions of the final network layer, i.e., the regression head, necessary to predict the parameter values of the assumed probability distribution, are presented and explained in detail in the subsequent paragraphs.

Confidence Estimation via Binary Classification

In the context of dense stereo matching, the task of predicting aleatoric uncertainty is commonly interpreted as confidence estimation (Tosi et al., 2018; Fu et al., 2019). Understanding confidence as the probability that a particular disparity estimate is correct, a CNN can be trained to distinguish between *correct* and *incorrect* disparity estimates, while using the class probability of *correct* as confidence score (see Sec. 3.2). To obtain this probability, the final layer of CVA-Net is defined as a 3D convolutional layer with a kernel size of $1 \times 1 \times 1$ followed by a sigmoid non-linearity (see Fig. 4.3a). In addition, a specific error definition is needed to cast the continuous range of disparity estimates into binary class labels used as ground truth for training and testing. For this purpose, the error metric proposed by Menze and Geiger (2015) is used: A disparity estimate d is assumed to be correct if either $|d - \hat{d}| < 3\,\text{pixels}$ or $|d - \hat{d}| < (\hat{d} \cdot 0.05)$, where $\hat{d}$ is the corresponding ground truth disparity. Using the predicted confidence score γ and the ground truth class labels $\hat{\gamma}$, this first variant of our network is trained by minimising the weighted binary cross-entropy loss:

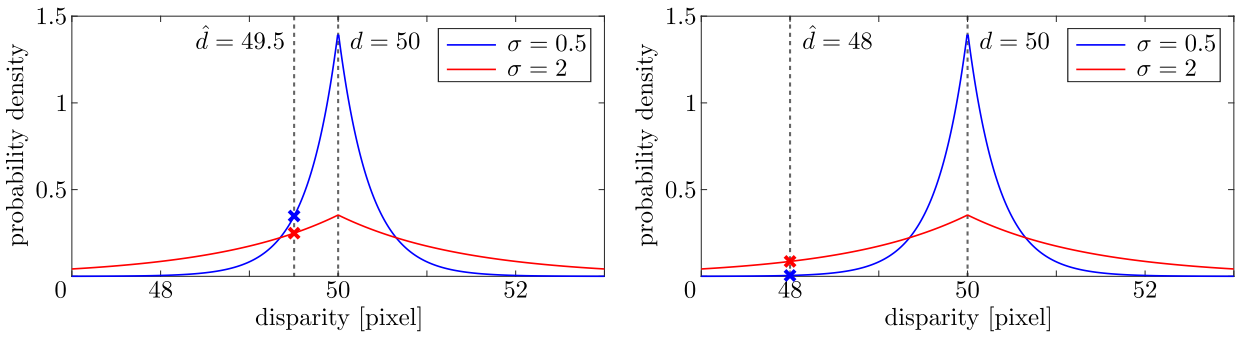
$$\begin{gathered} \mathcal{L}_{\text{Confidence}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} w_{\mathbf{p}} \cdot h(\gamma_{\mathbf{p}}, \hat{\gamma_{\mathbf{p}}})\,, \text{ with:} \\ w_{\mathbf{p}} = \hat{\gamma_{\mathbf{p}}} \cdot (w_{corr} - 1) + 1 \quad \text{and} \quad h(\gamma, \hat{\gamma}) = -\hat{\gamma} \cdot \log(\gamma) - (1 - \hat{\gamma}) \cdot \log(1 - \gamma)\,, \end{gathered} \tag{4.3}$$

where $\mathcal{D}$ is a set of pixels with known ground truth disparity. While the function h computes the standard binary cross-entropy, samples with a correct disparity estimate are weighted by the ratio between incorrect and correct training samples w_{corr}. This ratio is considered in the loss function to account for unbalanced training sets, which prevents the network from learning to preferably predict the more frequent class.

Uni-Modal Probabilistic Model

In contrast to the stochastic model based on confidence estimation, all subsequent variants learn to predict aleatoric uncertainty in a Bayesian way. For this purpose, first a specific probability distribution is assumed to describe the uncertainty contained in the data. The values of the parameters characterising this distribution are either predicted by CVA-Net, for example, the standard deviation representing the uncertainty, or derived from the cost volume using an operation defined by the stereo matching method the volume originates from, for example, the disparity estimate via the argmin operation. Note that the estimation of aleatoric uncertainty is assumed to be independent of the disparity estimation process. Thus, only the parameters representing the uncertainty are variable and adjusted during the training process, with the objective of maximising the likelihood of the corresponding ground truth disparity (Kendall and Gal, 2017), as illustrated in Figure 4.4. Following this procedure, aleatoric uncertainty can be learned as variance or standard deviation from the distribution of the disparity error, thus avoiding the need for a direct reference for the uncertainty, such as explicit parametrisations of the probability distribution.

Based on the common usage of the L1 norm in the context of training a CNN for the task of disparity regression (Mayer et al., 2016; Kendall et al., 2017b), the Laplace distribution is used to describe the aleatoric uncertainty, which in turn applies the L1 norm to the disparity residuals. To



(a) Small disparity error

(b) Larger disparity error

Figure 4.4: **Principle of likelihood maximisation in the context of aleatoric uncertainty estimation**. In both scenarios, the uncertainty is modelled as a Laplace distribution, using the estimated disparity d as mean and the predicted uncertainty as standard deviation σ to parameterise the distribution. (a) Having a small disparity error, here $|d - \hat{d}| = 0.5$ pixels, a smaller standard deviation is preferred, because it results in a higher likelihood for the reference disparity $\hat{d}$. (b) With a larger disparity error, however, also a larger standard deviation is preferable. Thus, uncertainty estimation can be learned based on the disparity error by maximising the likelihood of the ground truth disparity over the assumed probability distribution.

enable the use of common optimisers, the objective of the deep learning-based optimisation process is formulated as the negative log likelihood of this distribution:

$$-\log p(\hat{d}_{\mathbf{p}}|d_{\mathbf{p}}) \propto \frac{\sqrt{2}}{\sigma_{\mathbf{p}}}|d_{\mathbf{p}} - \hat{d}_{\mathbf{p}}| + \log(\sigma_{\mathbf{p}})\,, \tag{4.4}$$

where d is the estimated and $\hat{d}$ the ground truth disparity, while σ is the standard deviation of the assumed Laplace distribution representing the aleatoric uncertainty for the respective pixel. Similar to a procedure proposed in (Kendall and Gal, 2017) for using a loss function based on the negative log likelihood of a Gaussian distribution, σ is substituted with $s = \log(\sigma)$ in the loss function. This substitution makes the training process numerically more stable and prevents the loss function from being divided by zero. Thus, with this modification, CVA-Net is trained to predict the log standard deviation, using a 3D convolutional layer with a kernel size of $1 \times 1 \times 1$ without a subsequent non-linearity as final layer (see Fig. 4.3b). Finally, the loss function of the uni-modal probabilistic variant is defined as:

$$\mathcal{L}_{\text{Laplace}} = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{p}\in\mathcal{D}} \frac{\sqrt{2}}{\exp(s_{\mathbf{p}})}|d_{\mathbf{p}} - \hat{d}_{\mathbf{p}}| + s_{\mathbf{p}}\,. \tag{4.5}$$

Outlier-aware Mixture Model

The variant described before, models the uncertainty contained in the data using a Laplacian distribution and is thus based on the assumption that the disparity error of every pixel can be approximated reasonably well using such a uni-modal distribution. This assumption implies that a best match exists in the second image (having minimal matching cost) for every pixel in the reference image and that the probability of any other candidate to be the correct correspondence decreases

with increasing distance to this best match. Therefore, the usage of a uni-modal distribution is directly associated to the unique matching assumption introduced in Section 2.1.3. However, as already discussed in the context of this assumption, it is regularly violated, due to various reasons that result in either multiple or no potential matches. Consequently, also the usage of a uni-modal probability distribution to represent aleatoric uncertainty is not reasonable for all pixels in an image, leading to the need for a more sophisticated model.

With the mixture model presented in this section, this issue is addressed from the perspective of measurement reliability, explicitly taking care of cases which violate the aforementioned assumption. For this purpose, the approach of Vogiatzis and Hernández (2011) and Pizzoli et al. (2014) is adapted, which differentiates between two types of measurements produced by a depth sensor: good measurements that are uni-modally distributed around the correct depth and outlier measurement that are drawn from uniform distributions. Based on the considerations regarding the previous model, good measurements are again assumed to be Laplacian distributed. However, instead of demanding an exclusive representation by one of the two distributions, a weighted mixture of both distributions is assumed for every disparity estimate, following the procedure of Vogiatzis and Hernández (2011). In the context of this work and contrary to their approach, the parameters of this mixture distribution are not estimated via conventional convex optimisation, but in a Bayesian deep learning setup with the objective to maximise the likelihood, analogous to the optimisation based on the uni-modal variant described before. The loss function for this variant is thus defined as:

$$\mathcal{L}_{\text{Mixture}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} \alpha_{\mathbf{p}} \cdot \mathcal{L}_{\text{L}} + (1 - \alpha_{\mathbf{p}}) \cdot \mathcal{L}_{\text{U}} \,, \tag{4.6}$$

where the inlier probability α, together with the log standard deviations of the Laplace distribution s_{L} and the uniform distribution s_{U} are predicted by CVA-Net, again applying the substitution of $s = \log(\sigma)$. For this purpose, the regression head of this variant consists of three parallel 3D convolutional layers with kernel size $1 \times 1 \times 1$. One of these layer is followed by a sigmoid non-linearity to ensure that $\alpha \in [0, 1]$, while the output of the two other layers are directly used as log standard deviations to parameterise the assumed distributions (see Fig. 4.3c). The log likelihood terms of the two distributions themselves are further defined as:

$$\mathcal{L}_{\text{L}} = \frac{\sqrt{2}}{\exp(s_{\mathbf{p}})} |d_{\mathbf{p}} - \hat{d_{\mathbf{p}}}| + s_{\mathbf{p}} \,, \tag{4.7}$$

which is identical to the definition of the Laplacian distribution in the uni-modal variant and:

$$\mathcal{L}_{\text{U}} = \begin{cases} 0.5x^2 & \text{if } |x| \leq \gamma \\ \gamma|x| - 0.5\gamma^2 & \text{otherwise} \,, \end{cases} \tag{4.8}$$

which represents the log likelihood term of the uniform distribution. In this context, x is defined as the difference between the absolute disparity error and half the length of the interval with uniform distribution $r_{\mathbf{p}}$, resulting in: $x = |d_{\mathbf{p}} - \hat{d}_{\mathbf{p}}| - r_{\mathbf{p}}$. While the ground truth disparity $\hat{d}$ needs to lie in the interval $[d - r, d + r]$ to maximise the probability, x is minimised to prevent the network from predicting unreasonable large intervals. With the relationship between the interval length and the standard deviation σ_U of the uniform distribution, it further is: $r = \sqrt{3}\,\sigma_U$. The complete term $\mathcal{L}_{\text{U}}$ is set up in form of a Huber loss function (Huber, 1981), which combines the advantages of the

L1-loss (constant gradients for large values of x) and the L2-loss (less oscillation and thus more stability when x becomes close to zero). Finally, the variance of the presented mixture model is computed according to the general definition of mixture distributions:

$$\sigma^2 = \sum_{i=1}^{n} w_i(\sigma_i^2 + \mu_i^2 - \mu^2)\,, \tag{4.9}$$

where μ_i and σ_i^2 are the mean and variance of the ith component of the mixture distribution weighted by w_i and μ is the mean of the mixture distribution itself. Based on the assumption that the mixture distribution consists of only two components, a Laplace and a uniform distribution, and that the means of both components are equal to the estimated disparity, the computation of the variance σ_A^2 describing the aleatoric uncertainty can be simplified as follows:

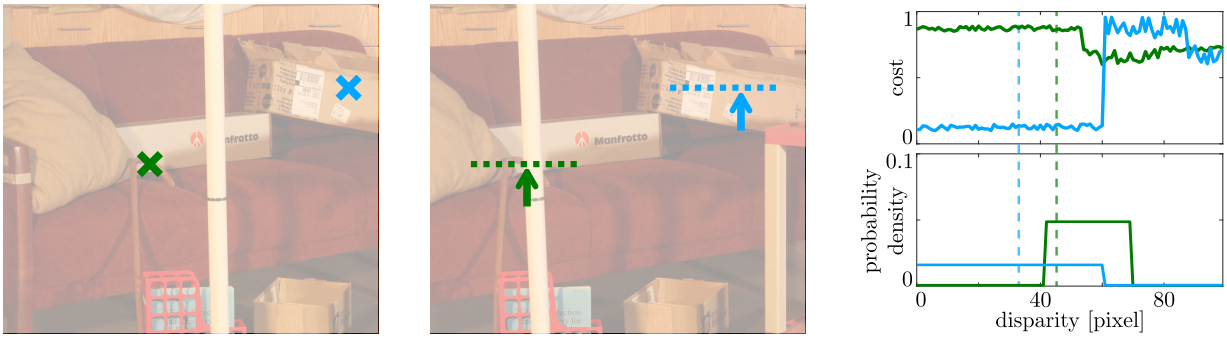
$$\sigma_A^2 = \alpha \cdot \sigma_{\mathrm{L}}^2 + (1-\alpha) \cdot \sigma_{\mathrm{U}}^2\,, \tag{4.10}$$

where σ_{L}^2 and σ_{U}^2 are the variances corresponding to the Laplace and the uniform component, respectively, and α is the inlier probability predicted by CVA-Net, as already defined before.

Scene-aware Model

The previously introduced outlier-aware model learns the ratio between inlier and outlier distribution completely from the training data, without posing any assumption on the composition of the resulting mixture distribution. In contrast, the last variant to be presented in this thesis links the composition of the mixture distribution with the characteristics of the depicted scene, resulting in a scene-aware model. In more detail, this variant is based on the discussion that arose from the unique matching assumption that pixels can be divided into two categories according to their compliance or violation of this assumption. As discussed in Section 2.1.3, such violations are commonly caused by scenarios that lead to multiple or no potential matches, for example, weakly textured areas or occlusions. Following this concept, in this variant the uncertainty assigned to a pixel is either modelled using a Laplace distribution, if a pixel in the reference image is expected to have an unambiguous correspondence in the second image, or with a uniform distribution, if the unique matching assumption is expected to be violated. Thus, the two types of distributions are not fused but used exclusively.

To be more specific, the two most common challenging scenarios in the context of dense stereo matching are considered in the definition of this model: weakly textured areas in an image and occluded regions. A typical behaviour for pixels in the reference image located in such weakly or non-textured areas is the occurrence of a wide and flat minimum in the corresponding cost curve (cf. Fig. 4.1d). Thus, multiple potentially matching pixels exist that are all characterised by similar costs and therefore have a similar probability of being the correct correspondence. This behaviour can be approximated with a uniform distribution on an interval that covers all disparities leading to pixels in the second image that are located in the same weakly textured area as the pixel of interest in the reference image, as illustrated in Figure 4.5. In the case of occlusion, in contrast, no correct correspondence exists for a certain pixel in the reference image. However, the appearance of the corresponding cost curve may be similar to the ones belonging to pixels located in weakly



(a) Left image (b) Right image (c) Cost / probability density

Figure 4.5: **Basic principle of the proposed scene-aware model**. While a Laplace distribution centred on the supposed optimal disparity is assumed for pixels with an unambiguous global minimum in the corresponding cost curve, such a minimum can typically not be determined for pixels that are located in a weakly textured area in the reference image (blue) or that are occluded in the second image (green). Instead, the error resulting from such scenarios is assumed to be uniformly distributed over a certain interval. The optimisation objective of the scene-aware model encourages CVA-Net to predict these intervals to be as small as possible, while still containing the correct disparities (indicated in (b) by arrows and in (c) by dashed lines).

textured areas, being characterised by the absence of a distinct global minimum. Due to the non-existence of a correspondence, the most basic approach would be to assume a uniform distribution over the whole disparity search range and thus to assign the same probability of being associated to the correct match to all possible disparities. However, the presence of occlusion requires an object in the scene that is located between the camera and the object point belonging to the pixel of interest in the reference image, hence occluding this point in the second image. Based on this considerations, the interval of uniform distribution can in principle be narrowed down to the pixels depicting this foreground object, as illustrated in Figure 4.5.

As errors arising from both, weakly textured and occluded regions, are assumed to be uniformly distributed in specific intervals, while the error of pixels fulfilling the unique matching assumption is expected to be Laplacian distributed, a loss function similar to the one of the previous variant is proposed:

$$\mathcal{L}_{\text{Scene}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} c_i \cdot \mathcal{L}_{\text{L}} + (1 - c_{\mathbf{p}}) \cdot \mathcal{L}_{\text{U}} + \beta_{\mathbf{p}} \cdot h(o_{\mathbf{p}}, \hat{o}_{\mathbf{p}}) \,, \tag{4.11}$$

where $\mathcal{L}_{\text{L}}$ and $\mathcal{L}_{\text{U}}$ are the Laplacian and the uniform likelihood terms as defined in Equations 4.7 and 4.8, respectively. However, as mentioned before, the two types of distributions are not mixed to describe the uncertainty corresponding to the disparity estimate of a certain pixel but used exclusively. Thus, c is a binary variable indicating whether the unique matching assumption is met or not. According to the definition of this binary classification discussed earlier, c is defined as: $c = \neg o \wedge \neg t$, where o specifies if the correspondence in the second image is occluded and t whether the pixel in the reference image is located in a weakly textured area. While t can be determined based on the reference image directly, for example, based on the criterion specified by Scharstein and Szeliski (2002), o is predicted by CVA-Net in addition to the log standard deviation (see Fig. 4.3d). In order to optimise the capability of predicting whether a pixel's correspondence is occluded or

not, the loss function is extended by a binary cross-entropy term h as defined in Equation 4.3, minimising the difference between the predicted occlusion values o and the corresponding ground truth $\hat{o}$. In this context, the pixel-dependent weight $\beta_{\mathbf{p}}$ is defined as:

$$\beta_{\mathbf{p}} = \beta_{\text{BCE}} \cdot (\hat{o}_{\mathbf{p}} \cdot (\beta_{\text{occluded}} - 1) + 1) \,. \tag{4.12}$$

It considers the class imbalance between non-occluded and occluded pixels using the ratio of their frequency in the training set as β_{occluded} as well as a static weight β_{BCE}, which is used to balance the influences of the binary cross-entropy term and the likelihood term.

4.3 Epistemic Uncertainty Estimation

In this second part of the methodology, the estimation of epistemic uncertainty is addressed. For this purpose, the cost volume construction $f^{(c)}$ and the subsequent disparity estimation $f^{(d)}$ are examined jointly (cf. Eq. 4.2). For this purpose, both functions are combined to estimate a disparity map $\mathbf{D}$ and an epistemic uncertainty map $\mathbf{U}_E$ from a stereo image pair $(\mathbf{I}_L, \mathbf{I}_R)$, which fulfils the assumptions stated in Section 4.1:

$$f^{(c,d)}(\mathbf{I}_L, \mathbf{I}_R) = f^{(d)}(f^{(c)}(\mathbf{I}_L, \mathbf{I}_R)) = (\mathbf{D}, \mathbf{U}_E) \,, \tag{4.13}$$

The basic idea of the approach presented in this section is to transform an existing CNN architecture, which is known to work well for the task of dense stereo matching, into a BNN, treating the convolutional layers of this architecture in a probabilistic manner. More specifically, the network parameters are sampled from a probability distribution which is optimised during training using variational inference, instead of treating these parameters as point estimates which are optimised directly. Similar to the previous section on aleatoric uncertainty, the functional and the stochastic model are again distinguished, being addressed in Sections 4.3.1 and 4.3.2, respectively. Note that aleatoric uncertainty is neglected for the moment, but integrated into the concept presented in this section in the following Section 4.4.

4.3.1 Functional Model

As became evident in the literature review, a wide range of approaches exist to estimate a disparity map from a planar rectified stereo image pair. With the aim of realising a method based on a BNN, a CNN-based approach that is trainable in an end-to-end manner is a natural choice, as this group allows for a relatively simple transformation with respect to the functional model and has demonstrated results superior to both, conventional hand-crafted methods as well as those that are partially based on machine learning. In addition, also the number of parameters of the network architecture $|\theta|$ serving as foundation is crucial, because the number of parameters of the variational distribution $|\phi|$ that are to be learned might be a multiple of $|\theta|$, depending on the stochastic model assumed. Thus, a probabilistic adaptation might be significantly larger than its deterministic baseline. In contrast, the size of the intermediate feature maps is commonly not affected by such a transformation into a BNN, resulting in only minor changes of the memory

footprint. Based on these considerations, it appears to be reasonable to select an architecture based on 3D rather than 2D convolutional layers, because of their often significantly lower number of parameters (cf. Sec. 3.1).

Specifically, the GC-Net architecture presented by Kendall et al. (2017b) is used as deterministic baseline in this thesis, which demonstrates good accuracy for the task of dense stereo matching, while having a relatively low number of parameters (∼2.8 Mio.), mainly justified by the absence of fully-connected layers (for details on this architecture, refer to Section 2.3.1). To transform this architecture into a probabilistic variant, the parameters of the network are no longer learned directly, as it is done by conventional deep learning and which would result in constant point estimates for every parameter, but sampled from a probability distribution which is defined by the stochastic model presented in the following section. In this context, the network parameters θ are sampled anew for every individual forward pass k, which results in slightly different variants of the same network $f_{\theta_k}^{(c,d)}$ and thus also in disparity estimates that vary with each sample:

$$f_{\theta_k}^{(c,d)}(\mathbf{I}_L, \mathbf{I}_R) = \mathbf{D}_k \,. \tag{4.14}$$

Carrying out several such forward passes, this procedure is commonly referred to as Monte Carlo sampling, whereas the employment of a trained BNN for testing with K Monte Carlo samples can be understood as sampling from an ensemble of K different neural networks. Thus, similar to other ensembling approaches (see Sec. 3.3), the disparity estimates resulting from several such samples k with $k \in \{1, .., K\}$ are combined, to compute the mean and variance of the distribution of these predictions:

$$\mathbf{D}(\mathbf{p}) = \bar{d}_{\mathbf{p}} = \frac{1}{K} \sum_{k=1}^{K} d_{\mathbf{p},k} \,, \tag{4.15}$$

$$\mathbf{U}_E(\mathbf{p}) = \frac{1}{K-1} \sum_{k=1}^{K} (d_{\mathbf{p},k} - \bar{d}_{\mathbf{p}})^2 \,. \tag{4.16}$$

Aggregating the resulting disparity estimates d of a pixel $\mathbf{p}$ over k samples, the average disparity estimate $\bar{d}$ and the variance σ_E^2 are used to obtain a disparity map $\mathbf{D}$ and an epistemic uncertainty map $\mathbf{U}_E$, respectively (cf. Eq. 4.13). This procedure is justified by the observation that deviations between different disparity estimates assigned to the same pixel reflect the model's uncertainty to determine the correct disparity, which allows to approximate the corresponding epistemic uncertainty based on these deviations.

Similar to the concepts presented by Zeng et al. (2018) and Brosse et al. (2020), not all parameters of GC-Net are treated in a probabilistic manner in the following, but some remain deterministic. Brosse et al. (2020) argue that it is sufficient to only model the final layer(s) of an architecture probabilistically in order to assess the epistemic uncertainty and to benefit from the positive effect of ensemble learning on the accuracy. Compared to a fully probabilistic approach, the proposed procedure minimises the growth of parameters to be learned as well as the computational overhead. However, both works only investigate such a setup in the context of classification. Due to the clear differences between architectures employed for classification tasks and those designed for dense stereo matching, in this work, a different approach is followed: Only the weights belonging to

convolutional filter kernels used in the feature extraction step (2D convolutions) and the multi-scale feature matching step in the encoder of the cost volume optimisation (3D convolutions) are treated probabilistically. In contrast, the parameters belonging to operations that are used to up-sample the intermediate feature maps (3D transposed convolutions), which is carried out in the decoder part of the cost volume optimisation step, are retained deterministically (cf. Fig. 2.4).

Besides the desired capability to estimate epistemic uncertainty, treating some parts of the network in a probabilistic manner further allows to reduce the model capacity without decreasing the accuracy of the estimated disparity maps. For this purpose, the number of filter channels n_c is adjusted, which is set to $n_c = 32$ for almost all layers of the original GC-Net architecture and to multiples of 32 if the spatial resolution of the feature maps is reduced in the inner layers of the encoder-decoder structure (cf. Sec. 2.3.1). As demonstrated by the results of preliminary experiments, n_c can be reduced by 25% to 24 channels without affecting the performance of the probabilistic variant and shows only a minor deterioration using 16 channels. The accuracy of the deterministic variant, however, is affected directly when reducing the number of feature channels. Such an adaptation of n_c does not only reduce the number of parameters of the network, which is already a positive effect itself especially with respect of the fact that treating the network in a probabilistic manner increases this number. Moreover, this adaptation of n_c also reduces the size of the intermediate feature maps as the feature channel number specifies one of the four dimensions of such a feature map as illustrated in Figure 2.3. Because the intermediate feature maps have the same size for both the deterministic and the probabilistic variant if the same number of filter channels is used, the ability to lower this number for the probabilistic variant results in a reduction of the memory footprint caused by the intermediate feature maps compared to the deterministic baseline. Finally, smaller intermediate feature maps also lead to less operations that have to be carried out to process them, which reduces the computational effort and thus minimises the computational overhead of training a BNN instead of its deterministic counterpart. In summary, the proposed transformation of the GC-Net architecture into a probabilistic variant using 24 filter channels increases the number of parameters to be learned only marginally from about 2.8 to 2.9 Mio. (assuming that the stochastic model is defined as described in the following section), while reducing the memory footprint of the intermediate feature maps by 25%.

4.3.2 Stochastic Model

To use the previously defined BNN for the purpose of Bayesian inference, the posterior distribution $p(\theta|\mathcal{D})$ of the network parameters θ given a set of training data $\mathcal{D}$ is required. However, as discussed in Section 2.3.2, computing and thus also sampling from this exact posterior distribution is typically an intractable problem. Therefore, in this work, the approach of variational inference is applied, aiming to learn the parameters ϕ of a variational distribution q that approximates the exact posterior distribution. To minimise the number of parameters to be learned as well as the computational overhead arising from VI compared to conventional deep learning, it is assume that the variational distribution over the latent variables, i.e., the network parameters, factorises as:

$$q(\theta_1, \theta_2, ..., \theta_n) = \prod_{i=1}^{n} q(\theta_i)\,. \tag{4.17}$$

This assumption is commonly referred to as mean field approximation, whereas a naive form is used in this work, assuming a partition into independent groups of single latent variables. The result is a diagonal Gaussian posterior, similar to the one proposed by Graves (2011). Consequently, the parameters of the variational distribution consist of a mean vector $\boldsymbol{\mu}$ and a diagonal variance-covariance matrix $\boldsymbol{\Sigma} = \mathbf{I} \cdot \boldsymbol{\sigma}^{\mathbf{2}}$, where $\mathbf{I}$ is the identity matrix, so that every network parameter treated in a probabilistic manner is drawn from an independent Gaussian distribution: $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. According to Graves (2011), this further allows to calculate the overall KL divergence between the exact posterior distribution and the variational distribution as the sum of the divergence terms corresponding to the individual partitions of the variational distribution:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^{n} KL(q_\phi(\theta_i)||p(\theta|\mathcal{D}))\,, \tag{4.18}$$

Using the KL divergence as regularisation term, while minimising the absolute difference between a disparity estimate d and the corresponding ground truth $\hat{d}$, the proposed BNN is trained end-to-end in a supervised manner using training data $\mathcal{D}$:

$$\mathcal{L}_{\mathrm{Epistemic}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} |d_{\mathbf{p}} - \hat{d}_{\mathbf{p}}| + \beta_{\mathrm{KL}} \cdot \mathcal{L}_{\mathrm{KL}}\,, \tag{4.19}$$

where β_{KL} is a hyper-parameter used to balance the two parts of the loss function. Relying on the concept of stochastic variational inference (Hoffman et al., 2013), the training procedure of the proposed probabilistic variant of GC-Net does not differ from the one used for the deterministic baseline in the sense that it can be easily integrated into an arbitrary deep learning framework and allows to apply common optimisation algorithms. However, as became evident in the literature review (cf. Sec. 3.3), stochastic neural networks commonly show a worse convergence behaviour compared to their deterministic counterparts, if the effects of the stochastic sampling of parameters during training are not considered. In this context, the major issue is the high variance of gradient estimates between consecutive training steps that arises from the stochastic nature of this sampling procedure. Using the same set of sampled parameters for all training examples in a mini-batch, i.e., all examples that are aggregated to perform a single back propagation and weight update, the gradients are correlated which prevents this variance from being minimised by averaging over the estimates of multiple examples in such a mini-batch (Wen et al., 2018). Thus, the basic idea of using mini-batches instead of single training examples to improve the convergence behaviour does not hold in this case. To mitigate this problem, Kingma et al. (2015) proposed the reparametrisation trick, which minimises the variance of gradients. While Flipout as proposed by Wen et al. (2018) requires more floating point operations during the weight update, it allows to reduce this variance even further, making it the means of choice in this work.

4.4 Joint Uncertainty Estimation

The third and last part of the methodology aims to fuse the concepts of aleatoric and epistemic uncertainty estimation presented in the previous sections, to obtain a consistent approach considering both sources of uncertainty in the context of dense stereo matching. The result is a realisation

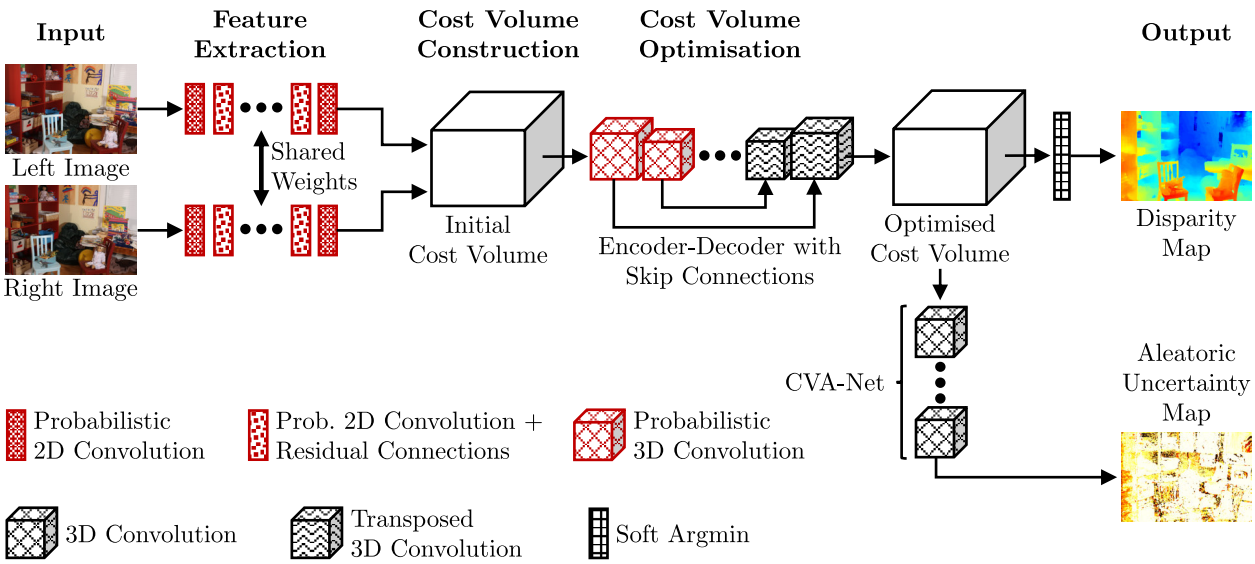


Figure 4.6: **Combination of the probabilistic variant of GC-Net and CVA-Net.** While the probabilistic adaptation of the GC-Net architecture is trained to predict a disparity map corresponding to the left image of a planar rectified stereo image pair, the probabilistic convolutional layers further allow to estimate the corresponding epistemic uncertainty via Monte Carlo sampling. CVA-Net is integrated as a separate branch, operating on the optimised cost volume to additionally predict an aleatoric uncertainty map. Source: Adapted from Mehltretter (2020).

of the functional relationship shown in Equation 4.1, estimating a disparity map $\mathbf{D}$ together with an uncertainty map $\mathbf{U}$ based on a planar rectified stereo image pair, which fulfils the assumptions stated in Section 4.1.

For this purpose, the functional models, i.e., the neural network architectures, of CVA-Net (see Sec. 4.2.1) and the presented probabilistic variant of GC-Net (see Sec. 4.3.1) are fused, as shown in Figure 4.6. While the basic definitions of both architectures remain unchanged, CVA-Net receives the whole optimised cost volume as input in this setup, instead of operating on a cost volume extract, which is possible due to the fully convolutional character of this CNN architecture. As a consequence of this integration, the result of CVA-Net is influenced by the stochastic nature of the early stages of the probabilistic GC-Net. More precisely, the estimated aleatoric uncertainty map directly depends on the network parameters randomly drawn from the posterior of the variational distribution. Thus, the uncertainty map varies for each Monte Carlo sample k, which makes it necessary to aggregate the uncertainty maps corresponding to all samples drawn to obtain a consistent result:

$$\mathbf{U}_A(\mathbf{p}) = \frac{1}{K} \sum_{k=1}^{K} \sigma^2_{A,\mathbf{p},k} \,, \tag{4.20}$$

where σ^2_A represents the aleatoric uncertainty computed according to one of the probabilistic models presented in Section 4.2.2. The computation of the disparity map $\mathbf{D}$ and the epistemic uncertainty map $\mathbf{U}_E$ remain equal to the definitions in Equations 4.15 and 4.16. Under the assumption that the aleatoric and the epistemic uncertainty are randomly and independently distributed, quadratic error propagation is applied to obtain the overall uncertainty associated with the disparity estimate

of a pixel $\mathbf{p}$:

$$\sigma_{\mathbf{p}}^2 = \sigma_{A,\mathbf{p}}^2 + \sigma_{E,\mathbf{p}}^2 \,, \tag{4.21}$$

from which the definition of the overall uncertainty map $\mathbf{U}$ follows as $\mathbf{U} = \mathbf{U}_A + \mathbf{U}_E$. Note that contrary to the exclusive training of CVA-Net, in this context, the disparity estimate is not fixed, but optimised jointly along with the aleatoric uncertainty. To achieve such a joint optimisation, the loss function of the probabilistic GC-Net (cf. 4.19) is adapted as follows:

$$\mathcal{L}_{\text{Full-Uncertainty}} = \mathcal{L}_{\text{Aleatoric}} + \beta_{\text{KL}} \cdot \mathcal{L}_{\text{KL}} \,, \tag{4.22}$$

where $\mathcal{L}_{\text{Aleatoric}}$ represents a loss function corresponding to one of the variants that allow to learn the prediction of aleatoric uncertainty in a Bayesian way (cf. Eq. 4.5, 4.6, 4.11) as presented in Section 4.2.2 and β_{KL} as well as $\mathcal{L}_{\text{KL}}$ are as defined in Equations 4.18 and 4.19. Consequently, the estimation of disparity and aleatoric uncertainty is learned together exploiting the principle of likelihood maximisation, while the estimation of epistemic uncertainty is further enabled by the usage of a BNN trained via VI.

4.5 Discussion

To close the chapter on the methodology, the approach and its components to estimate uncertainty in the context of dense stereo matching presented in the previous sections are analysed with respect to their advantages as well as the underlying assumptions and the corresponding limitations.

General Remarks

The basic assumption on the input data is that the stereo image pairs to be processed are captured simultaneously and are planar rectified. The assumption of rectified images holds as long as the relative orientation between the two image of a stereo pair is known from a prior calibration or can be accurately determined from image information, for example, based on feature points visible in both images that can be matched unambiguously. Note that the presented approach is able to tolerate minor deviations of a point correspondence in the second image from the related epipolar line due to the information of the local neighbourhood considered. However, such deviations are not explicitly modelled, leading to a decrease in accuracy for minor deviations and a potential failure for larger ones. On the other hand, an accurate time synchronisation of the stereo setup is a crucial precondition to guarantee simultaneously captured images in case of dynamic scenes, because temporal effects are not considered in the approach presented. Thus, changes of the scene depicted that occur in between the recording of the two images and that may be caused by moving entities would directly lead to incorrect depth estimates. Lastly, although this work focuses on the processing of terrestrial images, from a conceptual point of view, the presented approach can also be applied to other types, such as aerial images. However, it might be reasonable to adapt the definition of the cost volume to better reflect the probably rather different distribution of depth compared to terrestrial scenes. In addition, the increased computational effort and memory footprint caused by the typically much larger aerial images must also be considered.

CVA-Net Architecture

The network architecture CVA-Net developed in the context of this thesis allows to estimate aleatoric uncertainty from an arbitrary cost volume. More precisely, cost volumes with varying extent in the image and the disparity directions can be processed without the need of retraining, due to the fully-convolutional character of the network architecture presented and the employment of a global average pooling layer. Compared to other approaches working with cost volumes, CVA-Net does not need any pre-processing steps other than a normalisation, which avoids that potentially valuable information is eliminated. Moreover, cost volumes to be processed may be derived using different dense stereo matching procedures, making the employment of the presented approach highly flexible. However, this requires that a dense stereo matching method is not treated as a black box. Instead, intermediate results need to be accessible in order to be able to integrate CVA-Net, which might be a disadvantage compared to approaches that solely operate on disparity maps.

Having a closer look on the features extracted by CVA-Net, two limitations become evident: The receptive field is relatively small and only features defined on cost volumes are considered. While the size of the receptive field was determined empirically showing optimal results for the specified extent, it can be assumed that a broader context would be beneficial if incorporated in a more sophisticated way, for example, using a local-global fusion scheme as proposed by Tosi et al. (2018). However, it also has to be noted that the enlargement of the receptive field typically comes with an increase of the computational effort, especially noticeable due to the usage of 3D convolutions. Tightly coupled with this first point is the sole use of features from cost volumes, because not only a small receptive field may limit the amount of information available for the task of aleatoric uncertainty estimation, but also the neglect of additional modalities such as the RGB images or the disparity map. While both of these modalities might only provide limited further information in addition to those encoded in the corresponding cost volume, they could be well-suited to enlarge the receptive field due to their 2D dimensional nature. Note that both limitations, the size of the receptive field and the type of features considered, have been addressed in an initial investigation in (Heinrich and Mehltretter, 2021) which demonstrates that the consideration of these aspects further improves the quality of the uncertainty estimates obtained.

Aleatoric Uncertainty Models

To enable CVA-Net to estimate aleatoric uncertainty, two stochastic models are presented that are based on different mixture distributions optimised via likelihood maximisation. Contrary to the widely used approach of confidence estimation, the definition of the stochastic model in a Bayesian manner allows to quantify the aleatoric uncertainty in pixels or metric units and facilitates a natural combination with the epistemic uncertainty via error propagation. Moreover, the employment of a mixture instead of a uni-modal distribution enables the network to approximate the actual error distribution for a wider range of scenarios, whereas weakly textured and occluded regions of an image are explicitly taken into account. However, while the current definition covers two of the most challenging scenarios and thus frequent error sources, it does not allow to predict multi-modal

distributions. Thus, to accurately describe the error distributions related to pixels that belong to a repetitive texture or that are located close to depth discontinuities, and thus have two or more modes, require an extension of the approach presented, for example, using a Gaussian mixture model.

In addition, all stochastic models discussed in the context of this work treat the aleatoric uncertainty as being purely heterostedastic, neglecting homostedastic influences. Thus, all sources of uncertainty are assumed to influence each pixel differently and solely depend on the input. However, the presence of systemic effects that cause homostedastic aleatoric uncertainty is typically to be expected, for example, in form of uncertainty that arises from the camera calibration. Moreover, the models presented are limited to predict uncertainty and do not allow to incorporate prior information or to propagate uncertainty estimated in the context of preceding processing steps, such as the already mentioned camera calibration.

Bayesian Neural Network

In order to allow the estimation of epistemic uncertainty, an adaptation of the well-known GC-Net architecture is presented, in which the convolutional layers are replaced by their probabilistic counterparts and that is trained via VI. While the approach of drawing Monte Carlo samples at test time to determine the central moments of the posterior distribution is comparable to the concept of ensemble learning, the proposed employment of a BNN has the advantage that only the values of the variational parameters instead of the parameters of all trained models of an ensemble must be present when testing. Moreover, experiments have demonstrated that the capacity of the probabilistic variant can be reduced compared to the deterministic baseline without decreasing the accuracy of the resulting disparity maps. Thus, the memory footprint can be reduced and the computational overhead can be kept small. While the approach of transforming a CNN architecture that has proven to work well for the task of dense stereo matching in to a BNN is exclusively examined for the example of GC-Net in the scope of this work, the idea can also be applied to more recent architectures, which potentially further improves the accuracy of the disparity estimates.

In general, BNNs allow to specify prior distributions and to model dependencies between network parameters in form of correlations. Although the latter would be reasonable, especially in the context of a CNN which basically builds on the assumption that adjacent pixels are spatially related, it is not implemented in the context of the present work. The reason for the assumption of independent network parameters is the large amount of additional variational parameters that would need to be learned if correlations are considered. If, for example, a general mean field approximation instead of a naive one is assumed that allows to factorise the posterior variational distribution into groups of parameters belonging to individual layers, $n \times n$ covariances instead of n variances would need to be learned for a convolutional layer with n weights.

One of the drawbacks of stochastic neural networks is the need of Monte Carlo samples to approximate the moments of the posterior distribution. Thus, similar to ensembling procedures, multiple forward passes of a BNN have to be carried out at test time for a single sample of input

data, leading to a clearly higher computational effort compared to deterministic neural networks. While these forward passes are independent of each other and can thus in principle be computed in parallel, the large memory footprint and high computational effort of each individual forward pass currently prohibits such a parallel processing approach in practice, due to hardware limitations. Consequently, the estimation of epistemic uncertainty using a stochastic neural network and Monte Carlo samples typically increases the inference time significantly compared to a deterministic baseline.

Joint Uncertainty Estimation

The last part of the method presented addresses the combination of CVA-Net used to estimate aleatoric uncertainty and the probabilistic variant of GC-Net which allows to jointly estimate disparity and epistemic uncertainty. On the one hand, the concept followed in this work provides a modular view on the approaches presented, allowing to examine and employ them independently. On the other hand, both approaches can be seamlessly integrated, enabling end-to-end training which commonly leads to better results than optimising each component of a procedure on its own. In order to fuse the aleatoric and epistemic uncertainty estimated using the approach presented, it is proposed to apply quadratic error propagation, simply adding the variances representing both types of uncertainties. The basis for this procedure is the assumption that both uncertainties are randomly and independently distributed. While the assumption of independence is valid from a theoretical perspective - aleatoric uncertainty addresses the uncertainty inherent in the data, epistemic uncertainty the one inherent in a model - such a clear separation is commonly not achieved in practice. Often, both uncertainties rather influence each other, for example, because the model used to estimate the aleatoric uncertainty is also subject to epistemic uncertainty. Consequently, the assumption of independence of the uncertainties is a simplification that requires further investigations.

5 Experimental Setup

In this chapter, the experimental setup used to evaluate the methodology proposed in the previous chapter is presented. For this purpose, the objectives of the evaluation are introduced in Section 5.1, before presenting the datasets used for training and testing in Section 5.2. In Section 5.3, the framework for training the proposed approach is discussed, including an overview and a discussion of the hyper-parameter setting. This chapter closes with a presentation of the strategy and criteria for testing in Section 5.4.

5.1 Objectives

The overall objective of this work is the estimation of uncertainty in the context of dense stereo matching, which is addressed by an approach based on concepts of Bayesian deep learning. The quality of the estimated uncertainty and the effect on the disparity estimation capability itself can be evaluated using reference data. More precisely, the use of ground truth disparities allows a direct assessment of the accuracy of disparity estimates, while the uncertainty is evaluated based on the distribution of the disparity error. For this purpose, the results are evaluated in a quantitative manner based on the strategy and criteria presented in Section 5.4. The aim is to first examine the individual components independently, with the goal of enabling a more detailed analysis by mitigating effects from other parts of the approach. Subsequently, the approach is addressed as a whole, in order to investigate the overall performance and to analyse the integration of its components. Moreover, the following specific questions are investigated in the context of the experiments:

(1) How well is the proposed CVA-Net architecture suited to estimate aleatoric uncertainty? Is the use of features extracted from cost volumes sufficient to achieve accurate results? Is the proposed approach applicable to different dense stereo matching methods or do limitations exist that prevent such general validity?

This objective purely focuses on the functional model proposed in the context of aleatoric uncertainty estimation. Thus, only confidence estimation via binary classification is considered as stochastic model, being the most commonly employed approach. Using CVA-Net to estimate confidence, the results can be easily compared to those of state-of-the-art architectures presented in the literature, that address the same task. This allows to properly assess the result achievable with CVA-Net and to evaluate the suitability of cost volume-based features in comparison to other types of features. Lastly, the general validity of CVA-Net is investigated by carrying out the evaluation on cost volumes that correspond to different dense stereo matching methods, taking into account

conventional hand-crafted approaches with and without global optimisation as well as those based on deep learning, applied to images belonging to different datasets.

(2) How well suited are the probability distributions, implied by the different stochastic models used, to estimate the aleatoric uncertainty? Is it reasonable to make the kind of distribution assumed depend on the scene characteristics? Do specific circumstances exist leading to error distributions that can not be described well with the models proposed?

While the functional model is already evaluated by the investigations with respect to the first objective, CVA-Net is used as basis for the experiments related to the second objective. Keeping the functional model fixed, this procedure allows to focus on the effects that arise from the different stochastic models employed to estimate aleatoric uncertainty. Again, cost volumes that correspond to different dense stereo matching methods and images belonging to different datasets are examined to assess the strengths and limitations. In addition to the evaluation of the uncertainty maps as a whole, pixels belonging to typically challenging regions, such as weakly textured or occluded areas as well as those close to depth discontinuities, are investigated separately. This procedure allows to draw conclusions regarding the employed types of probability distributions and their combination, namely whether they are sufficient to describe the uncertainty in situations that commonly occurs in the context of dense stereo matching.

(3) What are the consequences of transforming a CNN developed for the task of dense stereo matching from a deterministic to a probabilistic formulation? Does this procedure influences the accuracy of the estimated disparities?

The main motivation for transforming a CNN architecture designed for the task of dense stereo matching into a stochastic neural network is the ability to estimate the epistemic uncertainty inherent in this model. However, the accuracy of the estimated disparities is not to be neglected, as it is desired to achieve at least the same accuracy as for the deterministic baseline. To validate this behaviour, both variants, the deterministic baseline and the probabilistic adaptation presented in this work, are tested on different datasets. Besides the comparison of the accuracy of the estimated disparities, further characteristics are examined: training and inference time, the memory footprint and the sensitivity towards domain gaps between training and test data. This extensive comparison supplements the discussion of the theoretical differences between both approaches, allowing to draw more specific conclusions on the suitability of a BNN for dense stereo matching.

(4) How good is the quality of the uncertainty estimates obtained by the overall approach? What is the contribution of the individual components to this results and how well do these components integrate? How well does the approach generalise to a wide range of datasets?

To address this last objective, the estimated uncertainty is evaluated with respect to the distribution of the disparity error, analysing the correlation and using sparsification plots (see Sec. 5.4). To obtain a better understanding of the individual components of the approach presented in this work, three different variants are examined: individual estimation of aleatoric and epistemic uncertainty and estimation of the combination of both uncertainties. Similar to the experiments related to the

other objectives, images from various datasets are used for testing to examine the general validity and to identify potential limitations of the approach presented.

5.2 Datasets

To allow for a comprehensive evaluation, for the experiments carried out in the context of this work, different datasets are used, which are briefly described in the following, considering synthetic as well as real indoor and outdoor scenes. All datasets consist of planar rectified stereo image pairs with known reference depth. An overview of the relevant characteristics of each dataset is given in Table 5.1, while examples for the reference images and the corresponding ground truth disparity maps are shown in Figure 5.1.

Sceneflow FlyingThings3D

The Sceneflow FlyingThings3D dataset presented by Mayer et al. (2016) is a synthetic dataset, showing scenes of various randomly located objects. While the depicted scenes are rather unrealistic, the large amount of data with available reference allows to train large neural network architectures in an end-to-end manner from scratch, which is often not possible using real data, due to the lack of ground truth data. Because of the synthetic nature of this dataset, ground truth disparities are available for all pixels of the reference images. However, the ground truth disparity maps contain some outliers with unreasonably small or large disparities in the range of [-2k;10k] pixels. To filter out these outliers, the 0.25% of samples with the smallest and largest ground truth disparities are discarded and not used for training or testing in the context of this work. After this filtering, the remaining 99.5% of samples have ground truth disparity values in the range of [0;191] pixels.

KITTI

The KITTI benchmark consist of datasets and challenges for various areas of research in the context of photogrammetry and computer vision and is one of the de facto standards for evaluation in many of these areas. The KITTI 2012 and 2015 stereo datasets presented in (Geiger et al., 2012) and (Menze and Geiger, 2015), respectively, contain real image pairs, which were captured using vehicle mounted stereo camera set-ups. These images show various street scenes from urban as well as rural environment in Germany, mainly captured under good weather conditions and with sufficient daylight. Challenges typically arise from specular reflections, overexposure, large distances between the cameras and objects and more generally from the complexity of the observed scenes. Based on LIDAR point clouds projected to the coordinate system of the reference image, ground truth disparity maps with disparities for about 30 % of the pixels are provided. In addition, for the samples of the KITTI 2015 datasets, the ground truth is improved by fitting 3D CAD models of cars into the initial point clouds. However, it is to be noted that the ground truth disparity maps do not contain any values for about the upper third of pixels, which is also visible in the example

Table 5.1: **Relevant characteristics of the datasets used in the experiments of this work.** Note that the number of samples does not specify the complete size of a dataset, but only considers samples with publicly available ground truth, which is required for both, supervised training and the evaluation of disparity and uncertainty based on reference data. Moreover, the values of the Middlebury dataset correspond to the quarter resolution images as provided by the authors.

	Image Resolution [px]	#Samples	Indoor / Outdoor	GT Density	GT Source	Max. GT Disp. [px]
Sceneflow	960×540	26760	synthetic	100%	synthetic	191*
KITTI	1240×376	394	outdoor	28%	lidar	232
Middlebury	up to 741×497	15	indoor	97%	structured light	200
InStereo2K	1080×860	2050	indoor	90%	structured light	328

*: Because the ground truth disparity values are located in the interval [-2k; 10k], the 0.25% of both, samples with the smallest and largest disparities, are discarded, reducing this interval to [0;191] for the remaining 99.5%.

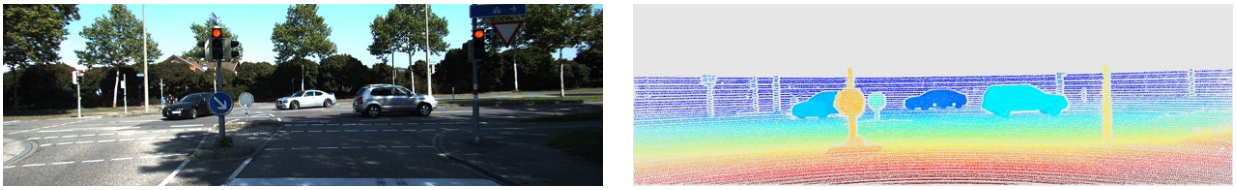
shown in Figure 5.1. In the context of the experiments of this work, both datasets, KITTI 2012 and 2015, are considered together and referred to as KITTI in the following.

Middlebury

The Middlebury stereo benchmark version 3, mainly containing samples from the Middlebury 2014 stereo dataset presented in (Scharstein et al., 2014), consisting of 15 stereo image pairs with known ground truth disparity, showing various indoor scenes. These scenes are captured with a static stereo set-up, varying the baseline and the image resolution for different samples. Dense and sub-pixel accurate ground truth disparity maps are obtained with a structured light-based approach. While the captured scenes are typically less complex compared to the previously introduced outdoor datasets, the more accurate ground truth allows for a more refined analysis, which is especially interesting with respect to the high accuracy achieved by recently presented approaches addressing the task of dense stereo matching. Moreover, also such indoor scenes may contain challenging scenarios, for instance large weakly textured areas that can be regularly observed on the floor or on walls. Note that in the experiments of this work, due to hardware limitations, the stereo images are processed in quarter resolution as provided by the authors.

InStereo2K

The InStereo2K dataset presented by Bao et al. (2020), also shows various indoor scenes, captured by a stereo set-up with structured light-based ground truth similar to the one used for the Middelbury dataset. Using two different but rather short baselines, the distance between the cameras and the scenes to be captured is kept small and the considered depth range is typically clearly smaller than the one of the Middlebury samples.



KITTI

Sceneflow Middlebury InStereo2K

Figure 5.1: **Examples from the datasets considered in the experiments of this work.** The figure contains one example per dataset, consisting of the reference image and the corresponding ground truth disparity map. Large disparities are shown in red, small ones in dark blue. Pixels without a ground truth disparity are shown in grey.

5.3 Training and Hyper-parameter Settings

In this section, the procedures used to train the approach investigated in this work as well as its components, namely CVA-Net and the probabilistic variant of GC-Net, are presented and the related hyper-parameter settings are described and justified. For this purpose, strategies and techniques applied to all variants are described first in Paragraph 5.3.1, before giving details of the individual training procedures in the subsequent paragraphs.

5.3.1 General Remarks

The approach presented in this work as well as its individual components are trained in a purely supervised manner, i.e., only data samples with known reference are considered in the computation of the loss. In addition, the disparity range considered during training is limited to [0,191] pixels, thus pixels with a ground truth disparity outside of this range are discarded and not used for training the network parameters. Consequently, the cost volumes, used as input to train different variants of CVA-Net and computed as intermediate results by the variants of GC-Net, have a depth

of 192. Limiting the disparity range in this way leads to training data and thus also to intermediate feature maps of constant size, which is beneficial for the training procedure from a practical point of view. Note that this only applies during training and does not limit the range of disparities that can be considered at test time, as described earlier. To determine the optimal number of training epochs for every model, an early stop strategy is applied: The training is stopped as soon as the validation loss does not improve in three consecutive epochs and the weights of the model with the lowest validation loss are then used for testing. Lastly, all deterministic convolutional and transposed convolutional layers are initialised using the Glorot normal initialiser (Glorot and Bengio, 2010), sometimes also referred to as Xavier normal initialiser.

5.3.2 CVA-Net

In this section, the training procedure of the CVA-Net architecture using the different stochastic models presented in Section 4.2.2 is described. The evaluation is carried out on cost volumes computed by three different stereo matching methods, namely Census-based block matching (Zabih and Woodfill, 1994), Census-based SGM (Hirschmuller, 2008) and MC-CNN fast (Zbontar and LeCun, 2016). Both Census-based methods use a support region size of 5×5 to compute the Census masks. While in the context of Census-based block matching and MC-CNN fast the cost volumes resulting from the cost computation step are used without applying a global optimisation procedure, in the context of SGM the cost volumes resulting from the global optimisation step are used. The penalties of the SGM smoothness term are determined empirically, using the values that minimise the mean absolute error of the disparity estimates on the training data. Because of the clear differences of these stereo matching methods, individual hyper-parameter settings are used in the experiments, which are described in the following together with the topic of cost volume normalisation.

Cost Volume Normalisation

As stated in (Kim et al., 2019b), cost curves highly depend on the utilised stereo matching approach and may vary largely for different approaches. First, different approaches may produce matching costs within different intervals. For Census-based approaches (Zabih and Woodfill, 1994), for example, matching costs are computed using the Hamming distance, resulting in costs c_{census} within the interval $\{c_{\text{census}} \in \mathbb{Z} \,|\, 0 \leq c_{\text{census}} \leq |\mathcal{A}|\}$, where $|\mathcal{A}|$ is the number of pixels considered as local neighbourhood. MC-CNN (Zbontar and LeCun, 2016), on the other hand, measures matching costs in the form of correlation, leading to results in the interval $[-1, 1]$. Additionally, not only the interval containing the results may vary, but also the characteristics of the computed cost curves, such as the curvature and the number of local optima. This characteristic depends on the cost computation method, as well as on the utilised optimisation approach, such as Semi-global matching (Hirschmuller, 2008), and its hyper-parameter values. The latter kind of variations are domain depended data variations, which can, for example, be considered by choosing a representative set of training data or by applying techniques to improve the network generalisation capabilities. The variations of the interval containing the matching costs and thus, of the network input data,

however, should be minimised in order to allow for a faster and more stable training process. For this purpose, in the present work min-max normalisation is applied: the theoretical boundaries of a matching approach's range of results are used to normalise the associated cost volumes in a preprocessing step. The result is a 3D tensor of real values in the range $[-1, 1]$, in which the characteristics of cost curves are preserved relative to each other.

Training Procedure

For the experiments carried out in the context of this work, several variants of the proposed CVA-Net are trained, considering all stochastic models presented and different dense stereo matching methods to compute the cost volumes used as input, as described above. Note that CVA-Net needs to be trained individually with data from every dense stereo matching method considered in the evaluation to achieve good results, because the generalisation capability between cost volumes corresponding to different methods is rather limited due to the highly varying characteristics of the contained cost curves. However, this does not limit the practical relevance of CVA-Net, because the need for retraining does not depend on the stereo images to be processed, but only on the dense stereo matching method utilised, which is commonly a design decision made in advance.

All variants of CVA-Net are trained on the first 20 image pairs of the KITTI 2012 dataset, which allows a fair comparison against state-of-the-art approaches to estimate aleatoric uncertainty presented in the literature, as these typically employ the same training strategy. Three additional image pairs are used for the purpose of validation. To generate individual training and validation samples from the cost volumes corresponding to theses image pairs, tensors of size $13 \times 13 \times 192$ are extracted from normalised cost volumes corresponding to the left image of each pair. Every extract is centred on a pixel with available ground truth disparity, resulting in more than 2.7 million training samples. 512 of such extracts are bundled to one mini-batch during training. In this context, an epoch is defined as completed once the network has seen all training samples. The Adam optimiser (Kingma and Ba, 2015) is employed to minimise the loss function defined by the respective stochastic model with a learning rate of 10^{-4}, setting the exponential decay rates for the moment estimates to their default values $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To enforce generalisation, dropout (Srivastava et al., 2014) is applied to the global average pooling layer (cf. Tab. 4.1) with a rate of 0.5.

All parameters mentioned so far apply for all variants of CVA-Net. The number of training epochs necessary for the model to converge, however, varies depending on the respective variant, as shown in Table 5.2a. The ratio between training samples with incorrect and correct disparity estimates w_{corr} used in the variant based on binary classification (see Eq. 4.3) and the ratio between occluded and non-occluded pixels β_{occluded} used in the scene-aware variant (see Eq. 4.12) are determined based on the frequency of these respective classes with respect to the estimated and ground truth disparity maps used for training. This results in the following ratios: $w_{corr} = 0.63$ for Census-based block matching, $w_{corr} = 0.11$ for Census-based SGM, $w_{corr} = 0.24$ for MC-CNN and $\beta_{\text{occluded}} = 20$. The parameter γ, which governs the transition between the two parts of the Huber loss in the mixture model-based variant (see Eq. 4.8), is set to one, as commonly done in the literature, for

Table 5.2: **Overview of all models trained for the experiments.** Variants that are not considered in the experiments are indicated by a hyphen, all others have assigned the respective number of training epochs.

Stochastic Model	Census	SGM	MC-CNN
Confidence	8	9	9
Laplace	14	-	11
Mixture	17	-	12
Scene-aware	16	-	17

(a) The CVA-Net architecture is trained with the different stochastic models described in this work (see Sec. 4.2.2) using cost volumes that correspond to stereo image pairs of the KITTI dataset and originate from Census-based block matching, Census-based Semi-global matching and MC-CNN.

Variant	Training Sceneflow	Fine-tune InStereo2K
Deterministic	16	24
Deterministic + CVA-Net	-	9
Probabilistic	24	47
Probabilistic + CVA-Net	-	10

(b) Four different variants of the GC-Net architecture are trained, namely the original deterministic variant, the proposed probabilistic adaptation and the combination of both with CVA-Net enabling aleatoric or combined uncertainty estimation, respectively. While the Sceneflow FlyingThings3D dataset is used for training, fine-tuning is carried out on the InStereo2K dataset.

example, by Girshick (2015). Also β_{BCE} is set to one, which weights the binary-cross entropy term equally relative to the likelihood term in the loss function of the scene-aware variant (see Eq. 4.12). Finally, the ground truth labels for pixels that are occluded or located in weakly textured areas used in the loss function of the scene-aware variant (see Eq. 4.11) are obtained as described in Section 5.4.4.

5.3.3 Probabilistic GC-Net

To train the proposed probabilistic variant of GC-Net as described in Section 4.3, setting the basic number of feature channels to $n_c = 24$ and using the loss function defined in Equation 4.19, 21 thousand stereo image pairs contained in the Sceneflow FlyingThings3D dataset and designated for training are used. For the purpose of validation, another 100 image pairs are considered. From these image pairs, random extracts of size 512×128 are cropped and fed to the network during training. The large difference between the width and height of a sample is justified by the importance of seeing all potential correspondences of a pixel along the horizontal epipolar line in the second image of a stereo pair for as many pixels as possible, while being restricted by hardware limitations in the number of pixels that are considered in a single sample and are thus processed together. Seeing all potential correspondences is required in order to avoid the introduction of an artificial bias in the distribution of disparities than can actually be observed in the training data. Using a batch size of 1, one such extract from every image pair is seen per epoch. The Gaussian distributions that form the variational distribution and from which the parameters of the probabilistic 2D and 3D convolutional layers are sampled as $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ (cf. Sec. 4.3.2), are initialised with $\mu = 0$ and $\sigma^2 = 1$. The hyper-parameter β_{KL}, which is used to weight the KL divergence relative to the term of the L1 norm in the loss function of the probabilistic variant of GC-Net (see Eq. 4.19), is not set statically, but adapted during the training process. More precisely, β_{KL} is set to zero for the first training epoch, allowing the optimisation process to focus on adapting the variational parameters

with the exclusive objective of minimising the disparity error in the beginning of the training procedure. In the following five epochs, β_{KL} is incremented by 0.2 per epoch, gradually increasing the regularisation effect of the KL divergence. In all consecutive epochs, β_{KL} is constantly set to one. Finally, RMSProb (Tieleman and Hinton, 2012) with a learning rate of 10^{-3} is employed for optimisation.

Besides the probabilistic variant of GC-Net, the original deterministic variant of GC-Net as presented by Kendall et al. (2017b) is trained and used as a baseline. To ensure a fair comparison, this baseline is trained in the exact same way as described above in the context of the probabilistic variant. After training on the synthetic Sceneflow FlyingThings3D dataset, fine-tuning is carried out for both variants using 1800 image pairs of the InStereo2K datasets, in order to optimise the network parameters to process images showing real-world scenes. This procedure will further allow to analyse the effects of domain gaps between training and test data in the subsequent experiments. The number of epochs carried out for training and fine-tuning the deterministic and the probabilistic variants, respectively, can be found in Table 5.2b.

5.3.4 Combined Approach

In addition to the deterministic and probabilistic variants of GC-Net, both variants are also trained as combination with CVA-Net as described in Section 4.4, allowing for aleatoric and joint aleatoric and epistemic uncertainty estimation, respectively. Instead of training from scratch, i.e., using a random initialisation for the network parameters as described before, the parameter values of the epochs with the smallest validation loss, obtained when exclusively training the deterministic or probabilistic GC-Net without considering CVA-Net, are used to initialise the GC-Net part of both variants, respectively. This approach does not only accelerate the training procedure, but also improves the convergence behaviour of the more complex combined model, using good initial values for a share of the parameters. In addition, due to the increased memory footprint of the combined model, the size of the image extracts used as training samples is reduced to 384×96 pixels, because of hardware limitations. An overview of all variants and the corresponding number of epochs required to train the individual variants is provided in Table 5.2b.

To train the combination of the probabilistic GC-Net with CVA-Net, the optimisation objective as defined in Equation 4.22 is applied, using the loss function of the scene-aware model (cf. 4.2.2) as aleatoric component. On the other hand, the combination of the deterministic GC-Net with CVA-Net is trained based on the loss function of the scene-aware model only, not taking into account the KL divergence. For both variants, the scene-aware loss is extended by a coefficient β_{error}, weighting the individual training samples based on the corresponding disparity error. This procedure is necessary, because the disparity error arising from the estimates predicted by GC-Net is not well distributed over the disparity range considered, but mainly concentrated around zero. While this is a desired behaviour in the context of dense stereo matching, it motivates CVA-Net to preferably predict small uncertainties, thus resulting in an effect comparable to the one of imbalanced classes in a classification setup. Consequently, also disparity estimates that have a high deviation from the corresponding reference are assigned a small uncertainty, which may prevent

them from being identified as erroneous. To overcome this limitation, the training samples are weighted according to their disparity error, differentiating between values in three different ranges: $\beta_{\text{error}} = 1.3$ for a disparity error smaller than one pixel, $\beta_{\text{error}} = 7.7$ for a disparity error in the range of $[1, 5)$ pixels and $\beta_{\text{error}} = 12.5$ for a disparity error larger or equal than five pixels. The individual values of β_{error} are determined based on the error distribution of the training samples using GC-Net without CVA-Net. While the basic idea of this procedure corresponds to the approach applied in the case of imbalanced classes in a classification setup, the sensitivity of the hyper-parameter values determined as well as the number of individual ranges considered require further investigations that are to be carried out in future work.

5.4 Evaluation Strategy and Criteria

In this section, the strategy and criteria applied for the purpose of evaluating the approach and its components presented in this work are described. To obtain quantitative results, the disparity, confidence and uncertainty errors are quantified by the metrics described in Sections 5.4.1, 5.4.2 and 5.4.3, respectively. If not otherwise specified in the description of the individual experiments, for the purpose of computing quantitative results, 100 random image pairs are used per dataset during testing (and all 15 image pairs in case of the Middlebury dataset) that have not been seen by the network, i.e., the training, validation and test sets are strictly separated. The disparity range considered in the experiments, and thus the depth of the cost volumes used as input for CVA-Net or computed as intermediate results of the different GC-Net variants, is adapted to each dataset based on the maximum ground truth disparity as specified in Table 5.1. More precisely, the disparity interval considered ranges from zero to the first number that is larger than the maximum disparity and can be divided by 32 without remainder (note that the latter condition is not required from an algorithmic point of view, but from the implementation side). Besides the computation of these metrics considering all pixels of an image, a focus is put on regions that are particularly challenging in the context of dense stereo matching, as described in Section 5.4.4. Lastly, the number of Monte Carlo samples used for the probabilistic variant of GC-Net as well as for the combined approach presented in this work is justified in Section 5.4.5.

5.4.1 Disparity Error Metrics

To express the error of disparity estimates quantitatively, three metrics are used in the context of this thesis: the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE) and the Pixel Error Rate (PER). Because all three of them determine the error of a disparity estimate d with respect to the respective ground truth disparity value $\hat{d}$, only pixels $\mathbf{p}$ with a known reference disparity are considered in the computations:

$$\text{MAE} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} |d_{\mathbf{p}} - \hat{d}_{\mathbf{p}}| \,, \tag{5.1}$$

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} (d_{\mathbf{p}} - \hat{d}_{\mathbf{p}})^2} \,, \tag{5.2}$$

$$\mathrm{PER}_\tau = \frac{|\{\hat{\mathbf{p}} \,|\, \hat{\mathbf{p}} \in \mathcal{D} \wedge |d_{\hat{\mathbf{p}}} - \hat{d}_{\hat{\mathbf{p}}}| > \tau\}|}{|\mathcal{D}|}\,, \tag{5.3}$$

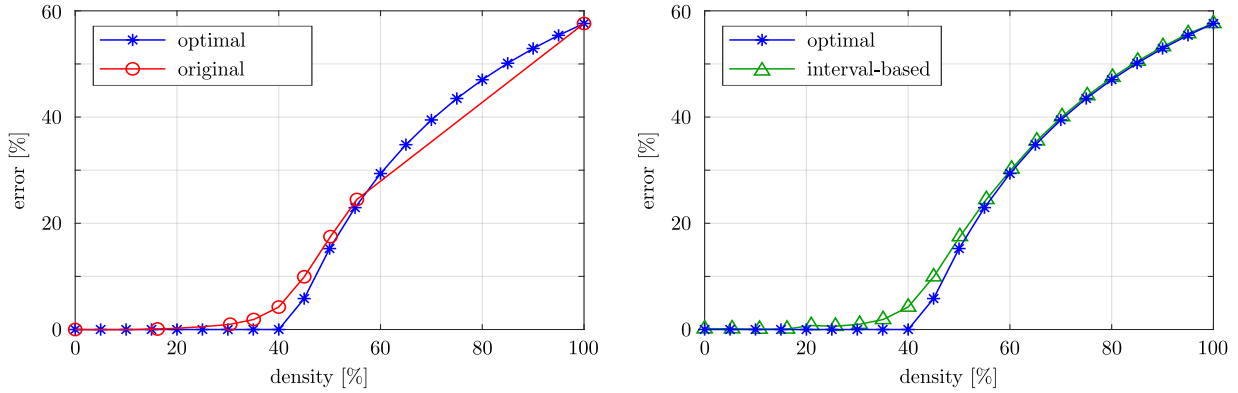
where τ specifies the threshold from which a disparity estimate is considered as erroneous, using one, three and five pixels as values for τ in the evaluation of this work. While the MAE weights all deviations equally and is typically easier to interpret, the RMSE is more affected by large errors and thus may indicate that an approach failed completely in a certain scenario. On the other hand, the PER allows for a finer analysis of the quantity of pixels that achieve a certain level of accuracy, specified by the threshold τ. In the subsequent evaluation, it is assumed that the data to be processed is free of gross errors. For this purpose, unreasonably small and large ground truth disparities are filtered out in advance for the Sceneflow dataset, as described in Section 5.2.

5.4.2 Confidence Error Metric

While the quality of the estimated disparity values is assessed with respect to reference data directly, such data do typically not exist for the uncertainty. Consequently, the different types of uncertainty estimates are evaluated with respect to the disparity error, following two different approaches for the assessment of confidence estimates and predicted standard deviations. For the purpose of evaluating the CVA-Net variant trained with binary classification as stochastic model (cf. Sec.4.2.2), a measure originally proposed by Hu and Mordohai (2012) is used. This measure relies on Receiver Operating Characteristic (ROC) curve analysis and is a well-established procedure in the field of confidence estimation. In this context, a ROC curve represents the error rate as a function of the percentage of pixels sampled from a disparity map in the order of increasing uncertainty. More precisely, in the first step, the 5 % of pixels having the lowest uncertainty are sampled from an estimated disparity map and the percentage of erroneous pixels in this set is determined. In the second step, this procedure is repeated, but using the 10 % of pixels with the lowest uncertainty. This procedure is further applied in 5 % steps, until the full density is reached and all pixels are considered, so that the error of the last set is equal to the overall error of the estimated disparity map. Of course, the usage of sampling distances different from 5 % would also be possible. In a final step, the Area Under the Curve (AUC) is computed for the estimated ROC curve, which is then used to express the accuracy of the uncertainty map regarding the detection of wrong disparity assignments in a single number. Assuming that an optimal uncertainty map contains smaller values for every correct disparity assignment than for any incorrect one, all pixels with a correct disparity assigned are sampled before any pixel with an incorrect one to obtain an optimal ROC curve (illustrated in Fig. 5.2) that defines the optimal AUC. Under this assumption, the optimal AUC can be computed directly from the overall error ϵ of a disparity map:

$$\begin{aligned} AUC_{opt} &= \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p} dp \\ &= \epsilon + (1-\epsilon)\ln(1-\epsilon)\,, \end{aligned} \tag{5.4}$$

where p is the percentage of pixels sampled from a disparity map. Because the optimal ROC curve corresponds to a perfect uncertainty map, any estimated uncertainty map results in a ROC curve that lies above or coincides with the optimal curve at every possible position and the closer the AUC of an uncertainty map is to the optimal value, the higher the accuracy.



(a) Original procedure

(b) The interval-based extension proposed

Figure 5.2: **Comparison of the two evaluation procedures based on ROC curve analysis.** (a) Following the original procedure, a ROC curve is not sampled within regions of equal uncertainty. This may result in an AUC with a high discretisation error and may cause segments of the approximated curve to lie below the theoretically optimal curve (here especially visible in the upper right part). (b) To minimise the discretisation error, additional sampling points are introduced within such regions. Ambiguities are avoided by estimating the error of these additional sample points based on the corresponding error intervals. Source: Mehltretter and Heipke (2021).

Following the described way of computing the AUC, pixels with equal uncertainty are sampled together during the approximation of the ROC curve to avoid ambiguous results. In consequence, the specified sampling distance may be exceeded significantly, if many pixels share the same uncertainty, resulting in a high discretisation error. While this was rarely the case for hand-crafted approaches and could therefore be neglected, it is a major issue for recent deep learning-based methods. Learning to predict aleatoric uncertainty via binary classification tends to push the predicted scores to the extreme values zero and one, resulting in large sets of pixels with the same uncertainty. Consequently, the described original procedure for sampling the ROC curve can lead to large distances in between two sample points (more than 40 % of all pixels) and, in extreme cases, the sampled ROC curve lies significantly below the theoretically optimal curve - a contradiction in terms. Both problems are exemplarily shown in Figure 5.2a.

As a solution to this problem, an interval-based extension to the previously described procedure is proposed in this work. The basic idea is to guarantee the compliance of the specified sampling distance for all segments of the ROC curve, and thus to minimise the resulting discretisation error. To achieve this objective, additional sampling points within sets of pixels with equal uncertainty are needed. However, since all pixels in such sets are assigned the same uncertainty, there is no unique order, which results in ambiguities when sampling points directly. To resolve these ambiguities, the developed interval-based extension considers two scenarios for each set of pixels with equal uncertainty, in which the pixels of a set are either sorted in ascending or descending order with respect to the absolute disparity error. While both scenarios allow for an unambiguous AUC computation respectively, an ascending order results in an approximation of the ROC curve with minimal AUC, a descending order maximises the AUC. Consequently, points sampled within regions of equal uncertainty are not assigned a unique error value, but an error interval defined by the minimal and maximal error resulting from the two different scenarios. To compare the

uncertainty estimated by different approaches in a quantitative way, the proposed interval-based extension is employed to sample additional points in regions of equal uncertainty, again in 5 % steps. To compute a unique ROC curve, the centre of a determined error interval is used as error value of the corresponding sample point. The computation of the AUC itself remains the same as in the original procedure. As shown in Figure 5.2b, the presented interval-based extension resolves the discussed problem of under-sampling inherent in the original procedure (cf. Fig. 5.2a).

5.4.3 Uncertainty Error Metric

A major drawback of the AUC metric is that it only considers the ratio of correct disparity estimates and the relative order of the estimated uncertainty values, while it neglects the actual magnitude of the estimated uncertainty and thus also the relation between a pixel's uncertainty estimate and its disparity error. While this behaviour is sufficient for the evaluation of confidence estimates, which represent the probability of a pixel's disparity estimate of being correct without providing any indication regarding the error magnitude, it does not cover all information provided by a standard deviation. To overcome this limitation, the correlation coefficient between the absolute disparity error and the estimated uncertainty in form of the standard deviation is used to evaluate the aleatoric, the epistemic and the joint uncertainty, estimated via the probabilistic uncertainty models presented in Section 4.2.2, the probabilistic variant of GC-Net (cf. 4.3.1) and the combined approach (cf. 4.4), respectively. In this context, the Pearson correlation coefficient is employed, using the following definition:

$$r_{\Delta d,\sigma} = \frac{\sum_{\mathbf{p}\in\mathcal{D}}(\Delta d_{\mathbf{p}} - \mu_{\Delta d})(\sigma_{\mathbf{p}} - \mu_{\sigma})}{\sqrt{\sum_{\mathbf{p}\in\mathcal{D}}(\Delta d_{\mathbf{p}} - \mu_{\Delta d})^2}\sqrt{\sum_{\mathbf{p}\in\mathcal{D}}(\sigma_{\mathbf{p}} - \mu_{\sigma})^2}}\,, \tag{5.5}$$

where $\Delta d_{\mathbf{p}}$ is the absolute difference between the estimated and the ground truth disparity of a pixel $\mathbf{p}$, σ is the estimated uncertainty in form of the standard deviation and $\mu_{\Delta d}$ as well as μ_{σ} are the mean disparity error and mean standard deviation, respectively. While the resulting correlation coefficient $r_{\Delta d,\sigma}$ always has a value between -1 and 1, the higher this value, the better the estimated uncertainty.

5.4.4 Region Masks

For the purpose of a more detailed analysis of the experimental results, besides the computation of the previously described metrics considering all pixels of an image, a focus is put on regions that are particularly challenging in the context of dense stereo matching, i.e., areas that are occluded in the second image of a stereo pair, that are weakly textured in the reference image or that are located close to depth discontinuities. To identify these image regions and to compute a binary mask for every type of region, the following definitions of Scharstein and Szeliski (2002) are used in the context of this thesis:

- Weakly textured regions: pixels for which the squared horizontal intensity gradient averaged over a 3×3 local neighbourhood is smaller than 4.

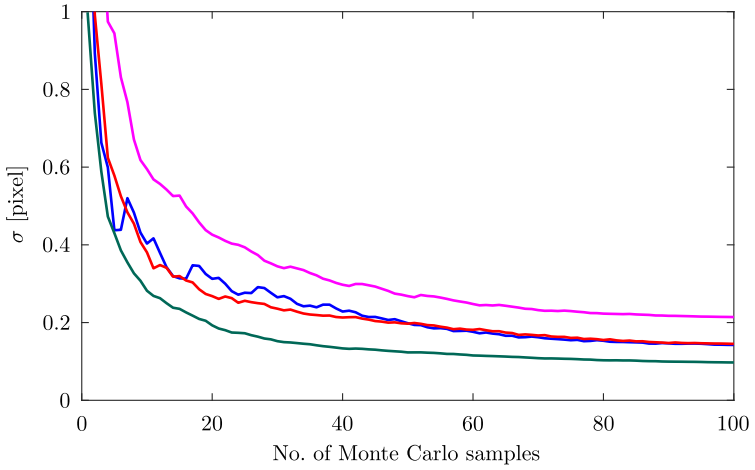


Figure 5.3: **Effect of the number of Monte Carlo samples drawn.** At test time, an average disparity estimate $\bar{d}$ is computed based on disparity estimates d_k with $k \in [1, .., K]$, corresponding to K different Monte Carlo samples (cf. Eq. 4.15). The larger the number of samples K considered in the computation of $\bar{d}$, the smaller the deviations between several independent computations of $\bar{d}$. Consequently, the standard deviation of such computations decreases with increasing K, converging against the epistemic uncertainty. This effect is shown for four different stereo image pairs, indicated by the four graphs, computing every disparity estimate $\bar{d}$ ten times for every number of Monte Carlo samples K examined. Source: Mehltretter (2020).

- Occluded areas: occlusions can be determined from ground truth disparity maps directly by projecting the disparity values from the reference image to the second image of a stereo pair. If the result is located outside the image plane or if it is projected to a pixel that is assigned a larger disparity further on, the respective pixel is set to one in the binary mask.

- Close to depth discontinuities: pixels for which the disparity differs by more than 2 pixels compared to the disparities of adjacent pixels, averaged over a 9×9 local neighbourhood.

Note that these definitions of occluded and weakly textured areas are also used in the context of training the scene-aware model for aleatoric uncertainty estimation (cf. Sec. 4.2.2) as described in Section 5.3.2.

5.4.5 Monte Carlo Sampling

For all experiments, the number of Monte Carlo samples drawn per stereo image pair to estimate the corresponding disparity and epistemic or joint uncertainty map is set to $K = 50$ (cf. Eq. 4.15, 4.16 and 4.20). As can be seen in Figure 5.3, an increasing number of samples improves the stability of the estimated disparity maps, in the sense that the deviations between multiple estimates computed based on the same stereo image pair decreases. However, at the same time, the computational effort grows linearly with respect to the number of Monte Carlo samples considered, thus leading to a longer runtime at test time. In consequence, the consideration of 50 Monte Carlo samples at test time seems to be a reasonable trade-off, since only minor improvements can be achieved when drawing additional samples.

6 Results and Discussion

In this chapter, the results of the experiments carried out in the context of this work are presented and analysed. For this purpose, this chapter is structured according to the evaluation objectives as described in Section 5.1: First, the CVA-Net architecture is evaluated in Section 6.1, before the stochastic models presented for the purpose of aleatoric uncertainty estimation are analysed in Section 6.2. In Section 6.3, the probabilistic variant of GC-Net is examined, focusing on the analysis of the differences to the deterministic baseline in Section 6.3.1, the quality of the estimated uncertainty in Section 6.3.2 and the influence of the KL divergence in Section 6.3.3. This chapter closes with a summary of the main findings and a discussion of the limitations in Section 6.4.

6.1 CVA-Net Architecture

In this first set of experiments, the functional model for the estimation of aleatoric uncertainty related to dense stereo matching presented in this work, the CVA-Net architecture, is evaluated. For this purpose, the performance of CVA-Net is compared against other state-of-the-art methods addressing the same task, namely CCNN (Poggi and Mattoccia, 2016c), LFN (Fu et al., 2019) and LGC-Net (using CCNN as local branch) (Tosi et al., 2018) which have been described in the literature review in Section 3.2, using cost volumes computed with Census-based block matching (Zabih and Woodfill, 1994), Census-based Semi-global matching (Hirschmuller, 2008) and MC-CNN fast (Zbontar and LeCun, 2016). Note that the implementations of the other confidence estimation approaches used in the evaluation are provided by the respective authors. As all of these methods realise this task solely using confidence estimation via binary classification as stochastic model, their results are compared against the respective variant of CVA-Net (cf. Sec. 4.2.2). The main difference between these methods and CVA-Net is the nature of the features utilised to predict the aleatoric uncertainty: While CVA-Net is based on features from cost volumes, the other methods learn to extract features from disparity maps only (CCNN and LGC-Net) or additionally from the RGB reference image (LFN). This difference in the features utilised also motivates the focus of this first set of experiments, analysing the advantages and drawbacks of uncertainty estimation via cost volume analysis.

Analysing the quantitative results presented in Table 6.1, it can be seen that the proposed CVA-Net architecture achieves results comparable to the state-of-the-art, outperforming it in some of the configurations evaluated. This is not only the case for the results on the KITTI dataset, on which CVA-Net was also trained on, but the proposed approach also shows state-of-the-art accuracy on the Middlebury dataset. This indicates that cost curves are characterised by similar features

Table 6.1: **Comparison against state-of-the-art confidence estimation methods.** The single entries show the theoretically optimal (*Opt.*) and the estimated AUC $\times 10^2$ averaged over 100 images of the KITTI dataset and all images of the Middlebury dataset (cf. Sec. 5.4). The smaller the values, the better, while *Opt.* is the best achievable value (cf. Sec. 5.4.2). Besides CVA-Net trained with binary classification as stochastic model, the confidence measures CCNN (Poggi and Mattoccia, 2016c), LFN (Fu et al., 2019) and LGC-Net (Tosi et al., 2018) are evaluated. This comparison is carried out using cost volumes computed with Census-based block matching (Zabih and Woodfill, 1994), Census-based Semi-global matching (Hirschmuller, 2008) and MC-CNN fast (Zbontar and LeCun, 2016).

avg. AUC $= 10^{-2}\times$	Opt.	CCNN	LFN	LGC-Net	CVA-Net
KITTI					
Census-BM	10.48	12.09	11.99	11.66	**11.17**
SGM	0.67	2.16	2.19	2.04	**1.90**
MC-CNN	2.11	2.89	2.90	2.74	**2.58**
Middlebury					
Census-BM	6.42	9.01	9.07	8.36	**8.19**
SGM	2.15	5.59	6.16	**5.30**	5.36
MC-CNN	3.38	5.20	5.27	**4.89**	4.91

independent of the dataset the evaluated stereo image pair belongs to. It furthermore shows that the concept of learning to estimate the uncertainty of a disparity assignment based on its cost curve generalises well over different datasets and is only marginally affected by a domain gap between training and test data.

Analysing the results in more detail, the proposed CVA-Net architecture reveals superior accuracy in particular in noisy regions of a disparity map. This applies regardless of whether these noisy disparity estimates belong to image regions that are weakly textured or characterised by high frequency patterns (cf. Fig. 6.1 and middle row of Fig. 6.2) or to areas which are occluded in one image of a stereo pair (cf. Fig. 6.3). Keeping in mind that all other methods estimate uncertainty based on the disparity map (and the reference image for LFN) only, it is evident that CVA-Net benefits from the additional information contained in cost volumes along the disparity axis. Methods that rely on the disparity map, often implicitly learn that areas with smooth disparity estimates have a high chance of being correct, which might be a wrong assumption, especially in image regions that are known to be challenging in the context of dense stereo matching, as demonstrated by the qualitative results. The corresponding cost curves, however, may reveal such ambiguities in the matching process, e.g., due to wide and flat global minima as illustrated in Figure 4.1, that indicate a higher uncertainty. The statement that cost volume-based information is beneficial is also supported by the fact that most disparity map-based approaches use a much wider receptive field. For example, LGC-Net uses a receptive field of size 48×48, while the one of CVA-Net has a size of 13×13 and is therefore provided with less information along the height and width axes of the image. Nevertheless, CVA-Net performs equally well and does show superior performance in some configurations.

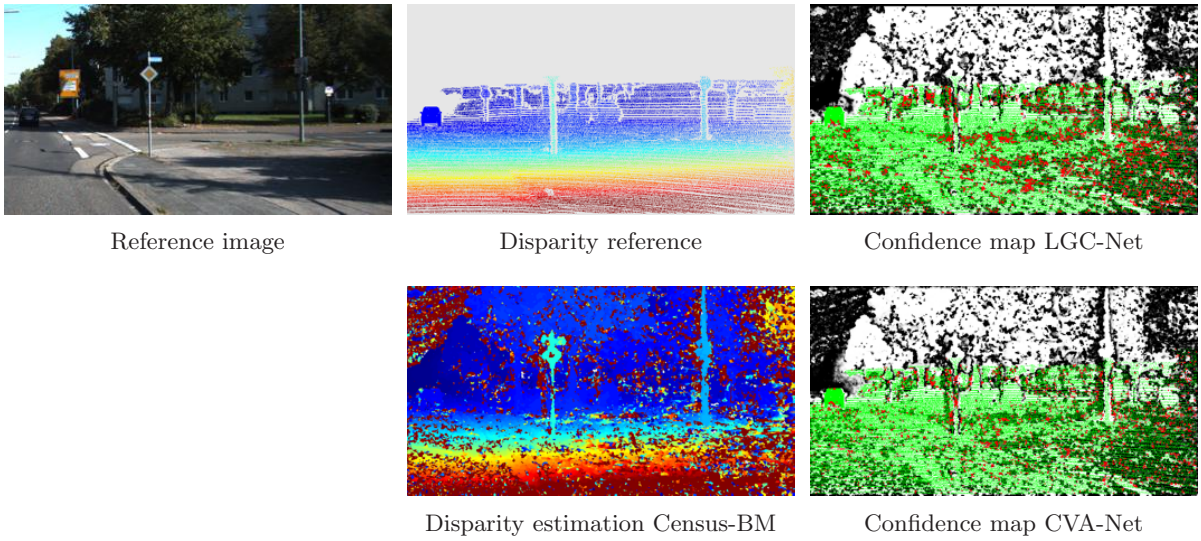


Figure 6.1: **Qualitative comparison on noisy disparity estimates.** The disparity maps show small values in dark blue to high values in red, pixels with unknown ground truth disparity are displayed in grey. In the confidence maps, a pixel's confidence is displayed from black (low) to white (high), while pixels with available ground truth disparity are displayed in colour. Green is assigned if either the assigned disparity is correct and the confidence c is larger than a threshold $\tau = 0.5$ or if the disparity assignment is wrong and $c \leq \tau$. Red pixels, on the other hand, indicate an incorrect confidence estimation. The confidence maps demonstrate that the presented cost volume-based approach is superior in image regions with noisy disparity estimates, for example, caused by weakly textured areas in the reference image, such as the street shown in the lower left corner or in the shadowed area on the right side.

On the other hand, depth discontinuities, which are also commonly identified as especially challenging in the context of dense stereo matching, demonstrate one of the limitations of the presented CVA-Net architecture. As can be seen in Figure 6.2, disparity estimates close to depth discontinuities are often assigned an incorrect confidence, regardless of whether the disparity map-based LGC-Net or the cost volume-based CVA-Net is used. It can be assumed that this is at least partly due to the fact that depth discontinuities and other associated effects, such as foreground-fattening, are not properly represented in the training set. The KITTI dataset, which was used for training to ensure comparability with the other confidence estimation methods, only provides sparse ground truth disparities and all methods are trained considering pixels with known ground truth only. Consequently, edges in disparity space caused by such depth discontinuities are rarely seen by the network during training. The results further indicate that CVA-Net typically suffers more from this problem than the other methods, often resulting in a wider margin of incorrect confidence estimates along a depth discontinuity, which is clearly visible in the example shown in the first row of Figure 6.2. This problem is probably caused by the relatively small receptive field of CVA-Net, which does not allow to consider a wider context, making it challenging to decide whether a pixel belongs to fore- or background. In this scenario, the additional information of the cost curves does often not support making this decision, but commonly simply indicates a high uncertainty. The reason for this behaviour can be found in the definition of the local neighbourhood used to compute the matching cost, which is specified as rectangular with constant extent for most dense

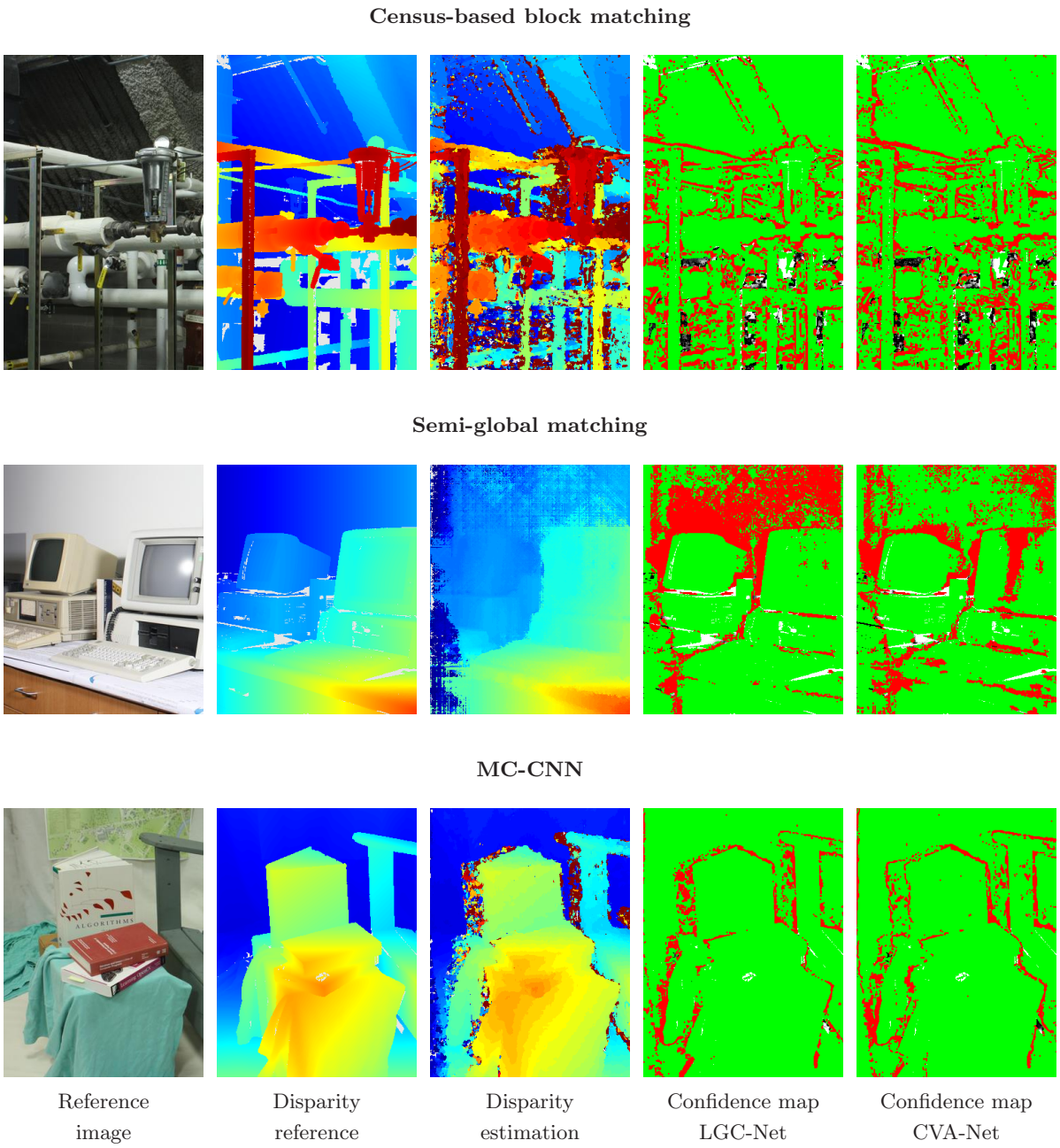


Figure 6.2: **Qualitative comparison in case of depth discontinuities.** Depth discontinuities are particularly challenging for all dense stereo matching methods considered, as can be seen in the respective disparity maps, and pose a common error case for all confidence estimation methods. This is especially visible if a dense disparity reference is available, such as for the samples of the Middlebury dataset shown in this figure. (For details on the colour coding, refer to Fig. 6.1).

stereo matching methods. As a consequence, these local neighbourhoods typically contain pixels from the fore- and the background when processing pixels close to depth discontinuities, which leads to at least two minima in the corresponding cost curve representing a depth in the fore- and the background, respectively. To resolve this problem, it may be beneficial to extend CVA-Net with a local-global structure similar to LGC-Net and to additionally take into account the information contained in the reference image, assuming that image gradients coincide with depth discontinuities (see Sec. 2.1.3). However, both proposals need further investigations and are subject of future work.

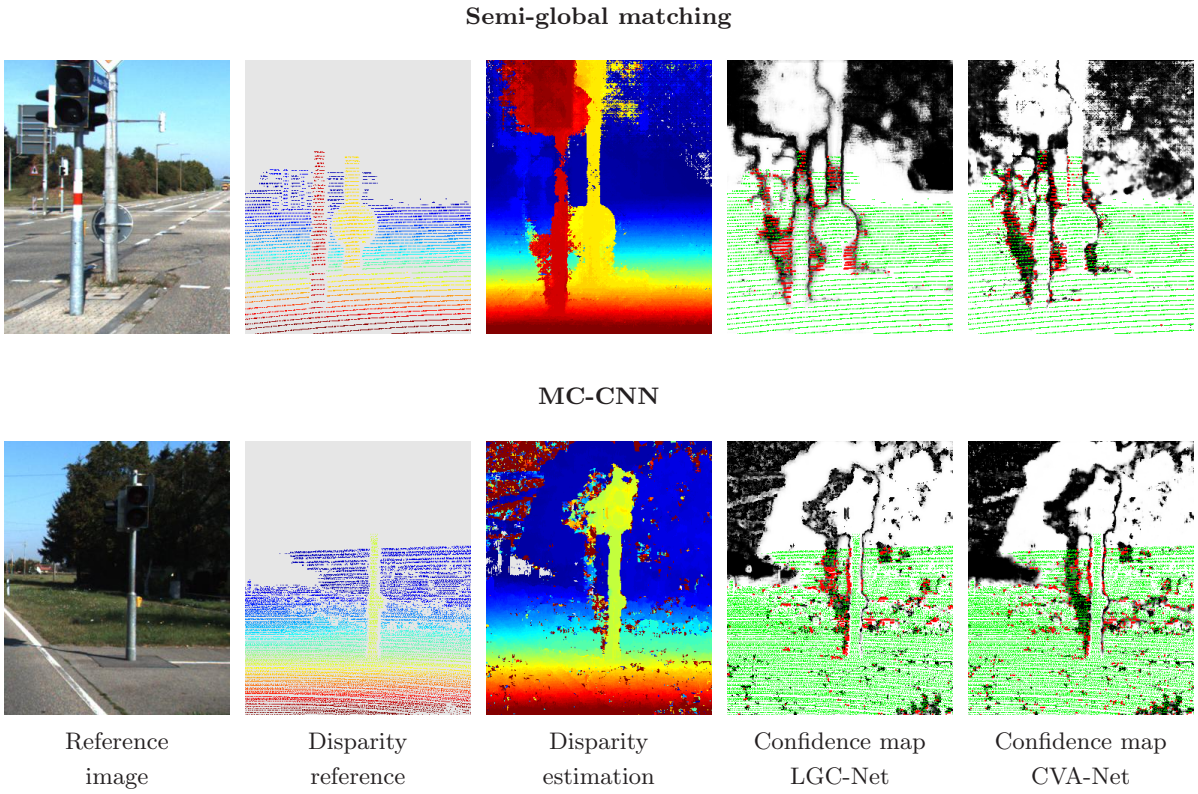
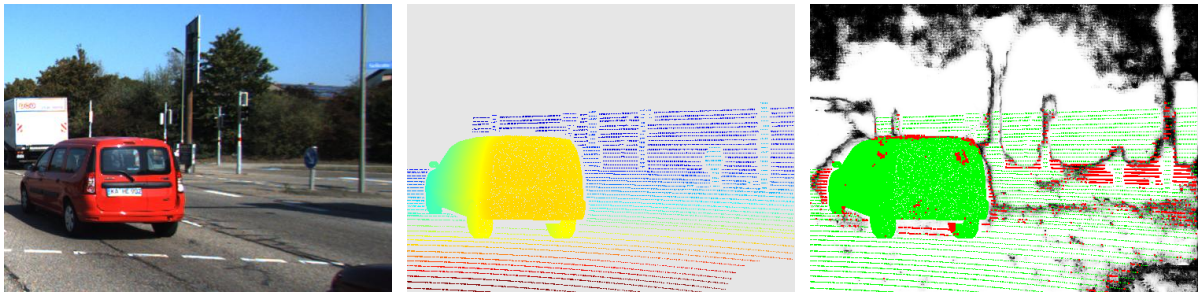
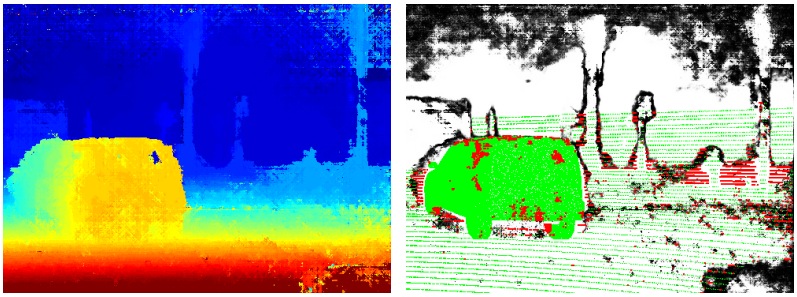


Figure 6.3: **Qualitative comparison in case of occlusion.** Disparity map-based approaches tend to assign high confidences to areas which are piece-wise smooth in the disparity map, even if the assigned disparities are not correct, e.g., caused by occlusions as can be seen to the left of the poles in both examples. CVA-Net, on the other hand, is able to detect these cases and predicts the correct confidence, because occlusion can typically be identified by analysing a pixel's cost curve. (For details on the colour coding, refer to Fig. 6.1).

Finally, the quantitative results presented in Table 6.1 demonstrate that the proposed CVA-Net architecture can successfully be applied to different dense stereo matching methods, covering a conventional local (Census-BM) and global approach (SGM) as well as a deep learning-based procedure (MC-CNN). However, the results also reveal that estimating the confidence corresponding to disparity estimates resulting from SGM is clearly more challenging than for Census-BM and MC-CNN, as can be seen by the significantly larger margin between the optimal AUC and the AUC corresponding to the estimations of SGM. Note that this is not only the case for CVA-Net, but also applies to the other confidence estimation methods evaluated. The reason for this problem is illustrated in Figure 6.4: While SGM is beneficial to minimise noisy disparity estimates and to correct local outliers, it may result in overly smooth disparity maps that suffer from the effect of foreground-fattening and the elimination of fine details, caused by over-smoothing depth discontinuities that are located close to each other. The latter is visible in the example especially between the poles on the right side of the image and between the rear wheels of the depicted car. Disparity map-based approaches often mistakenly consider such overly smooth disparity estimates as an indication for the correctness of these estimates, because these approaches often learn to associate the smoothness of disparity estimates with their correctness, as already discussed earlier. Also the additional information contained in the corresponding cost volume does typically not



Reference image | Disparity reference | Confidence map LGC-Net



Disparity estimation SGM | Confidence map CVA-Net

Figure 6.4: **Visualisation of the problem with SGM-based disparity maps.** Disparity maps resulting from SGM are characterised by smooth and less noisy estimates compared to the two other dense stereo matching methods considered in this section. However, SGM tends to over-smooth fine details and intensifies the problem of foreground-fattening, as can be seen between the rare wheels of the car and around the poles on the right side of the image. (For details on the colour coding, refer to Fig. 6.1).

allow to disclose this mistake, because SGM directly operates on the cost information, which leads to smoother cost curves that are characterised by a distinct global minimum in most cases. While this behaviour improves the accuracy of the disparity estimation, it distorts the conclusions drawn from the cost curve characteristics with respect to the associated uncertainty, for example, because originally occurring ambiguities are cancelled out which artificially reduces the assessed uncertainty (cf. Fig. 4.1). A potential solution to this issue may be to estimate the uncertainty of the initial cost volume originating from the underlying cost computation approach (e.g. Census, as in the experiments presented) and to propagate this uncertainty information through the global optimisation process of SGM, e.g., based on the approach presented in (Schönberger et al., 2018).

In conclusion, this first set of experiments demonstrates the ability of the proposed CVA-Net architecture to serve as functional model for the estimation of aleatoric uncertainty. Compared to state-of-the-art approaches from the literature addressing the same task, CVA-Net shows comparable performance and is superior in some scenarios. Especially the ability to assign correct confidences in the presence of occlusions or in weakly texture areas is a clear strength of the approach presented. Limitations, for example, in the context of depth discontinuities, are often not exclusive for CVA-Net, but can also be observed for the other methods evaluated, although CVA-Net appears to be more sensitive to this particular problem. Despite the fact that the SGM-related behaviour requires further investigation, the results still show that the proposed CVA-Net is applicable to cost volumes of quite different stereo matching approaches.

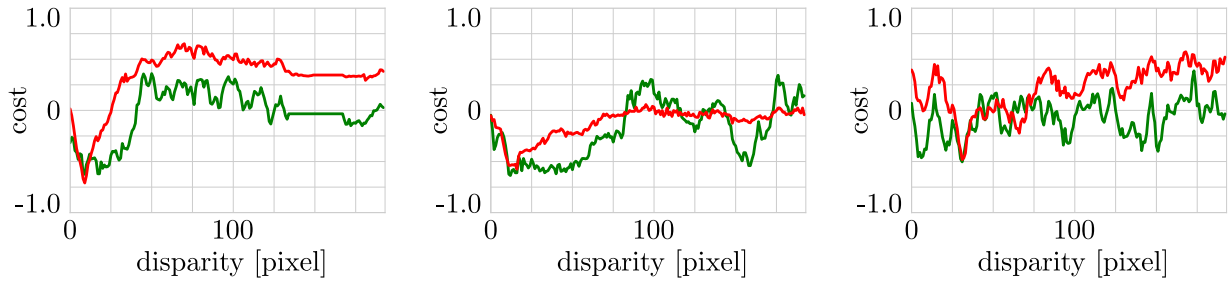


Figure 6.5: **Influence of Semi-global matching on the characteristics of cost curves.** Each figure shows the cost curves computed with Census-based block matching and Census-based Semi-global matching in green and red, respectively, for a particular pixel. All three examples show that SGM has a smoothing effect on the curves and almost always results in a curve with a clear and distinct global minimum, eliminating potentially occurring ambiguities.

6.2 Aleatoric Uncertainty Models

In this second set of experiments, the stochastic models proposed to learn the estimation of aleatoric uncertainty in a Bayesian way, as described in Section 4.2.2, are evaluated. Due to the missing possibility to evaluate confidence scores in terms of the error magnitude, the variant employing binary classification to learn the prediction of confidence is not considered in the evaluation carried out in this section. To ensure that differences in the results are actually caused by the varying stochastic models, the functional model is predominantly kept unchanged. More precisely, the CVA-Net architecture is used for this purpose, only adapting the final part of the architecture, the regression head, to allow for the prediction of the type and number of parameters needed by the respective stochastic model (see Fig. 4.3). As for the first set of experiments, cost volumes computed with different dense stereo matching methods are considered in the evaluation, namely Census-based block matching (Zabih and Woodfill, 1994) and MC-CNN fast (Zbontar and LeCun, 2016). Due to the remaining open problems with Semi-global matching discussed in the previous section, cost volumes corresponding to SGM are not considered in the experiments evaluated in this section.

Analysing the quantitative results shown in Table 6.2, it is evident that the mixture as well as the scene-aware model outperform the Laplacian model in all configurations evaluated, often by a clear margin. This behaviour is also visible in the illustration of the distributions of the predicted uncertainty with respect to the actual disparity error shown in Figure 6.6: While the Laplace model results in a more dispersed distribution, especially the scene-aware model leads to a clearly improved correlation. These observations already allow to draw a first conclusion, namely that the additional consideration of a uniform distribution in the uncertainty model is beneficial in assessing the actual error distribution. In addition, it can be stated that this modification not only improves the results in regions it was meant to, namely regions that are especially challenging in the context of dense stereo matching and thus often have a higher number of large errors (e.g. visible in Fig. 6.8c), but supports the estimation of aleatoric uncertainty in general. This effect can be explained by the implicit compromise when using the Laplace distribution only, namely the joint optimisation for small uncertainty estimates in areas where the correct disparity can be identified

Table 6.2: **Comparison of the stochastic models used for aleatoric uncertainty estimation.** The single entries show the Pearson correlation coefficients (cf. Eq. 5.5) between the disparity error and the standard deviations estimated with the individual models, considering all pixels (total) or evaluating specific image regions, namely weakly textured and occluded areas and pixels that are located close to depth discontinuities (cf. Sec. 5.4.4). For this purpose, 100 images of the KITTI dataset and all images of the Middlebury dataset (cf. Sec. 5.4) are considered, using cost volumes computed with Census-based block matching (Zabih and Woodfill, 1994) and MC-CNN fast (Zbontar and LeCun, 2016).

	Total	Weakly Textured	Occluded
Laplace	0.85	0.83	0.82
Mixture	0.88	0.85	0.85
Scene-aware	0.91	0.89	0.88

(a) Census-BM on KITTI

	Total	Weakly Textured	Occluded	Depth Discon.
Laplace	0.82	0.82	0.76	0.82
Mixture	0.86	0.84	0.75	0.84
Scene-aware	0.87	0.86	0.77	0.87

(b) Census-BM on Middlebury

	Total	Weakly Textured	Occluded
Laplace	0.86	0.84	0.84
Mixture	0.86	0.84	0.85
Scene-aware	0.89	0.87	0.87

(c) MC-CNN on KITTI

	Total	Weakly Textured	Occluded	Depth Discon.
Laplace	0.81	0.81	0.69	0.79
Mixture	0.81	0.81	0.69	0.79
Scene-aware	0.82	0.82	0.70	0.81

(d) MC-CNN on Middlebury

and localised well and very large uncertainties in areas where this is not possible. Treating these two situations individually, the consideration of a uniform distribution in challenging areas supports the estimation of uncertainty via a Laplace distribution in the other areas.

Examining the quantitative results in more detail, it becomes evident that the domain gap between the KITTI dataset the different models were trained on and the Middlbury dataset used for testing leads to a noticeable decrease in accuracy. This is especially visible for occluded regions, for which the correlation coefficient drops sharply for both dense stereo matching methods. In general, this behaviour can be explained by the different allocation of pixels to the individual regions: While the majority of pixels do not belong to challenging regions for the KITTI dataset, most pixels are either located in weakly textured or occluded regions for the Middlebury dataset, as illustrated by the region mask in Figure 6.8b. As these regions contain by definition less information valuable for both, dense stereo matching as well as the quantification of the associated uncertainty, it is to be expected that the quality of the estimated uncertainty decreases. While the presence and the extent of weakly textured regions can be determined from the reference image directly and are typically also clearly identifiable from cost curves as wide and flat global minima, occlusions do generally not lead to a global minimum in the cost curve at the position of the correct disparity (for example visible in Fig. 4.5c). In consequence, it is not only more challenging to identify the correct correspondence for a pixel located in an occluded region, but also to actually detect that a pixel belongs to such a region and to quantify the associated uncertainty. In this context, it is to be noted that for the Middlebury dataset, not only the percentage of occluded pixels is increased, but also the extent of individual occlusions, which further aggravates the previously described problem.

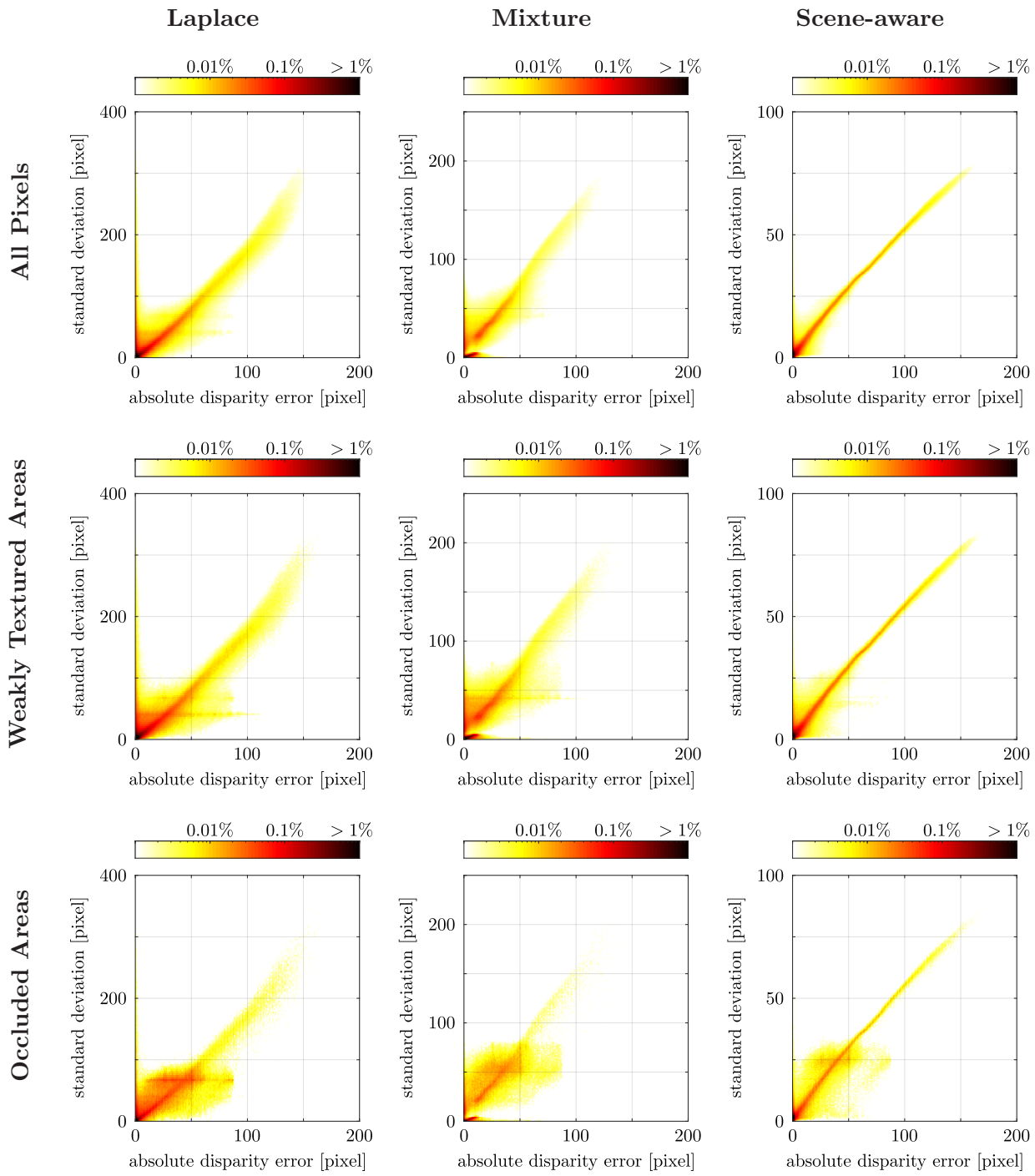


Figure 6.6: **Distribution of the predicted aleatoric uncertainty corresponding to Census-BM on KITTI.** The figure shows the aleatoric uncertainty predicted with the three different stochastic models with respect to the absolute disparity error computed with Census-based block matching in a logarithmic colour scale. While the visualisations in the first row consider all pixels of the evaluated KITTI images, the visualisations in the second and third row focus on especially challenging image regions, namely weakly textured and occluded areas, respectively.

This problem is also clearly visible in Figures 6.6 and 6.7, in which especially the distributions corresponding to occluded areas show a horizontal line that is more pronounced for the Middlebury dataset. This line indicates that all three models are unsure regarding the actual uncertainty,

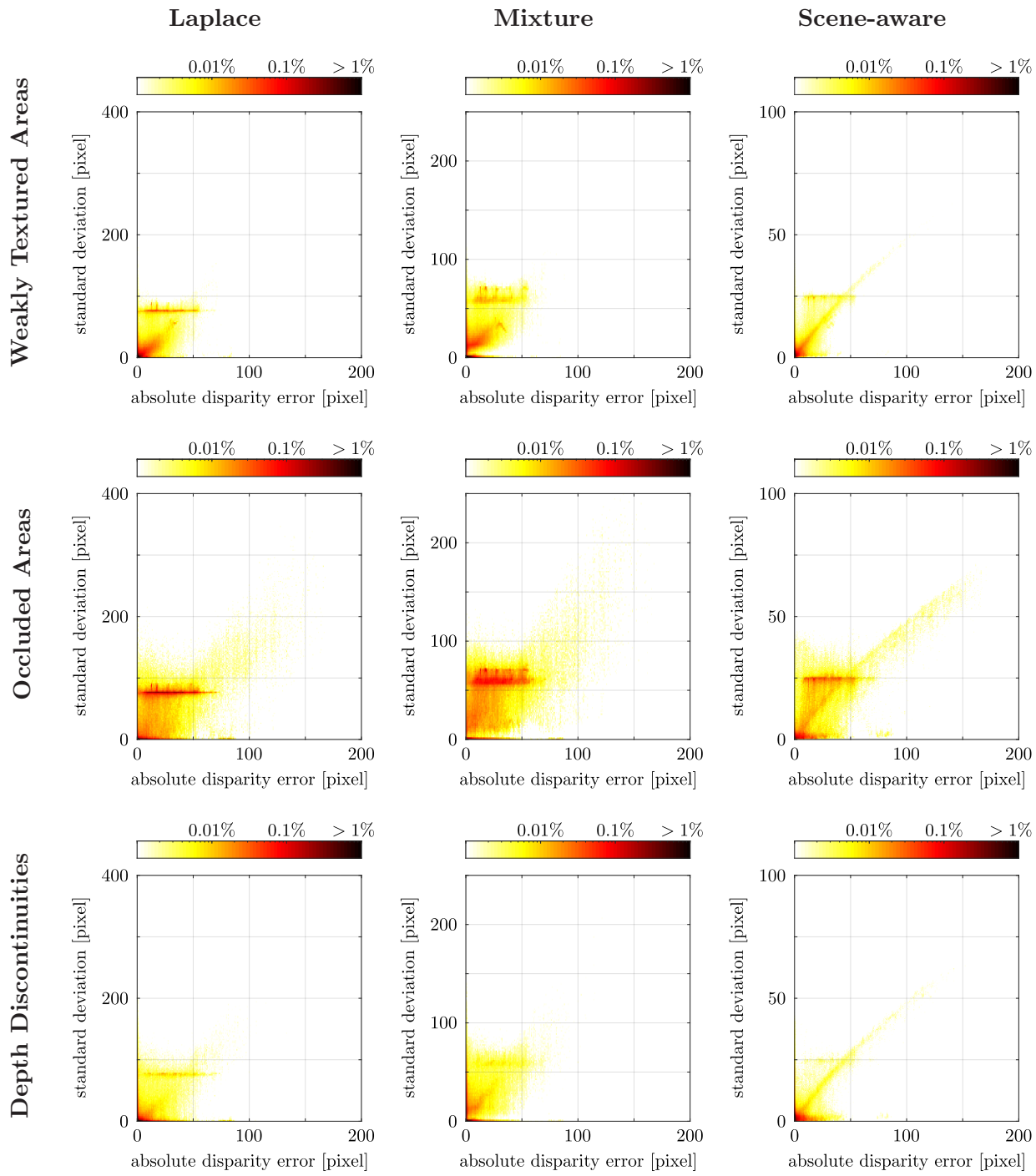
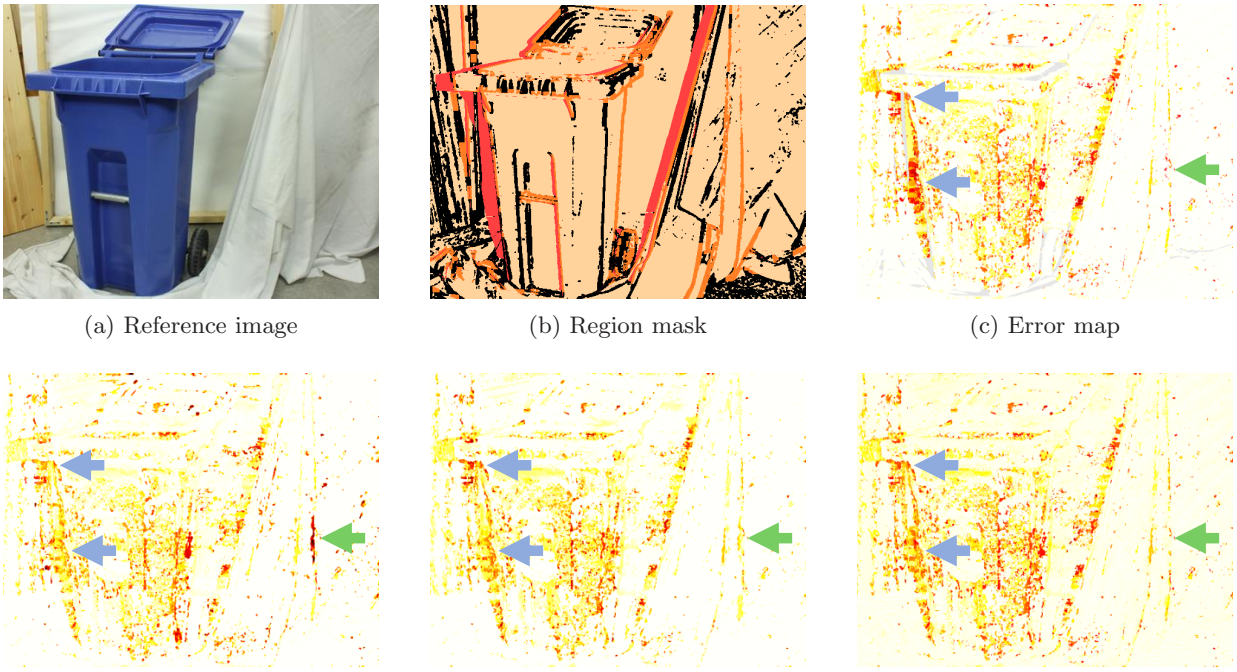


Figure 6.7: **Distribution of the predicted aleatoric uncertainty corresponding to MC-CNN on Middlebury.** The figure shows the aleatoric uncertainty predicted with the three different stochastic models with respect to the absolute disparity error computed with MC-CNN on the Middlebury dataset in a logarithmic colour scale. The three rows show the behaviour of these models in different challenging image regions, namely weakly textured areas, occluded areas and pixels located close to depth discontinuities.

predicting the same relatively large value for a wide range of disparity errors. However, both figures also demonstrate that the proposed scene-aware model is capable to mitigate this effect, which is also visible in the qualitative results shown in Figure 6.8 in which differences with respect to



(a) Reference image (b) Region mask (c) Error map

(d) Uncertainty map Laplace (e) Uncertainty map Mixture (f) Uncertainty map Scene-aware

Figure 6.8: **Qualitative results of the three evaluated aleatoric uncertainty models.** The figure shows the absolute disparity error in comparison to the uncertainty maps resulting from the three stochastic models used to estimate aleatoric uncertainty in a Bayesian way. The disparity map is computed with Census-based block matching on an image from the Middlebury dataset. The region mask highlights regions that are especially challenging in the context of dense stereo matching, showing weakly textured areas in beige, occluded areas in red and pixels close to depth discontinuities in orange. The error map and the uncertainty maps show small values in white and large ones in dark red / black. Note that the values of the three uncertainty maps are scaled to the same interval to allow for an easier comparison. The arrows indicate regions for which the uni-modal Laplace model under-estimates the error in occluded regions and over-estimates the error at depth discontinuities in blue and green, respectively.

occluded regions are highlighted by blue arrows. Moreover, the clearly smaller differences between the correlation coefficients belonging to the three aleatoric uncertainty models for MC-CNN fast on Middlebury (see Tab. 6.2d) compared to the other configurations evaluated are noticeable. This behaviour can be explained by the limited possibility of describing a more complex relation between the absolute disparity error and the estimated uncertainty with a single number. As can be seen in Figure 6.7, the three aleatoric uncertainty models result in clearly distinguishable distributions for this configuration, for which the scene-aware model shows the strongest relationship between the absolute disparity error and the estimated uncertainty.

After it could already be stated that the additional consideration of a uniform distribution is beneficial to estimate aleatoric uncertainty, clear differences in the results can also be observed with respect to the approach followed to combine the uniform and the Laplace distribution in the stochastic model. As described in Section 4.2.2, the mixture model estimates the uncertainty of each pixel as a weighted aggregation of these two types of distributions, where the weighting is also predicted by the network without any constraints or assumptions being placed on that weighting.

In contrast, the scene-aware model utilises only one type of distribution per pixel, with the decision which type is used being directly tied to the type of region a pixel belongs to. Comparing only the results of these two models, it can be seen that the scene-aware model outperforms the mixture model in all configurations. Thus, it can be concluded that the model assumption imposed by the scene-aware variant supports the training process and is more suitable for the problem addressed compared to the purely data-driven approach followed by the mixture model to determine the weighting. In this context, it is to be noted that for weakly textured regions the decision whether a Laplace or a uniform distribution is used to model the uncertainty of a pixel is derived from the image directly. On the other hand, this decision is based on a prediction for occluded regions, because the information whether a pixel is occluded is in general unknown at test time. While an accuracy of about 90% for both, the KITTI and the Middlebury dataset, implies that classification into occluded or non-occluded regions works rather well, precision and recall values of 0.5 and 0.6, respectively, show that there is still space for improvements. A particular challenge is the highly unbalanced distribution across the two classes, with occlusions being extremely underrepresented in the training data at less than 5%.

Lastly the focus is put on a region type that was not addresses in the discussion so far: depth discontinuities. In the definition of the scene-aware model as presented in this thesis, pixels that are located close to such depth discontinuities are not treated as belonging to a challenging region and the associated uncertainty is thus modelled using a Laplace distribution. This decision was made, because the cost curve corresponding to a pixel that belongs to such a region is typically not characterised by the absence of a clear minimum, but rather by the presence of multiple peaks at the disparities belonging to the depth of the objects that overlap in the image and form a depth discontinuity. While a Laplace distribution is not capable of capturing such a multi-modal characteristic, it is still better suited than a uniform distribution. However, as demonstrated by the quantitative results shown in Table 6.2 and the distributions of the uncertainty with respect to the disparity error shown in Figure 6.7, all three stochastic models are nevertheless able to assess the uncertainty of pixels located close to depth discontinuities reasonably well, achieving correlations comparable to the respective ones taking into account all pixels. Thus, as for the comparison of the overall correlations, the scene-aware model also demonstrates superior results for this particular region (cf. area marked by a green arrow in Fig. 6.8). Despite these good results also for depth discontinuities, further investigations should be carried out in the future, examining the possibilities to represent the uncertainty in such regions with multi-modal distributions, for example, via Gaussian mixture models. Based on the fact that such a model would better fit to the actual error distribution, this adaptation promises to further improve the results. However, in this context, it must also be noted that training a model that explicitly considers depth discontinuities requires a dense reference for the disparity to actually be able to determine ground truth labels indicating these regions.

6.3 Dense Stereo Matching using a Bayesian Neural Network

In the third and last set of experiments, the characteristics of the probabilistic variant of GC-Net presented in this work are analysed. For this purpose, in Section 6.3.1, this probabilistic variant is first compared against the deterministic baseline. In Section 6.3.2, the estimated uncertainty is examined, comparing the combined model against variants that account for either aleatoric or epistemic uncertainty only. Lastly, in Section 6.3.3, the influence of the KL divergence and the specified prior for the variational distribution is investigated.

6.3.1 Comparison to the Deterministic Baseline

In order to investigate the effects caused by the transformation of a deterministic CNN into a probabilistic one, in this section, the proposed probabilistic variant of GC-Net (in the following referred to as GC-Net$_{\text{prob}}$) is compared against the deterministic baseline as presented by Kendall et al. (2017b) (in the following referred to as GC-Net$_{\text{det}}$). In this context, the focus is put on the accuracy of the disparity estimated by these two variants, making sure that the objective of being able to assess epistemic uncertainty does not hinder the ability of disparity estimation. Furthermore, also practical aspects, namely the training and inference time as well as the memory footprint during training, are addressed.

Analysing the quantitative results presented in Table 6.3, it can be seen that in most configurations the MAE and the RMSE are lower for GC-Net$_{\text{prob}}$ compared to GC-Net$_{\text{det}}$, often by a clear margin. In this context, the difference between the two variants is significantly larger if training is carried out only on the synthetic Sceneflow dataset instead of also fine-tuning the network parameters with respect to the real-world InStereo2K dataset. In the latter case, the average disparity error becomes similar for both variants on all datasets but the one used for training. This indicates that the deterministic variant is more sensitive to the domain gap between synthetic training data and real-world test data than the probabilistic counter-part. On the other hand, GC-Net$_{\text{prob}}$ seems to discard detailed dataset specific knowledge more easily, as implied by the clearly stronger deterioration with respect to all metrics when being fine-tuned on a dataset different than the one used for the initial training, as can be seen by the results on the Sceneflow dataset comparing Table 6.3a and Table 6.3b.

In contrast to the improvements in the MAE and RMSE values, the results of GC-Net$_{\text{prob}}$ tend to contain more smaller errors shown by the often higher pixel error rate, especially evident if a threshold of one pixel is used. Combining these two observations, having a lower average error while producing more smaller errors, it can be concluded that the disparities estimated by GC-Net$_{\text{prob}}$ contain a lower number of large errors. This statement is also supported by the qualitative results shown in Figure 6.9. While the error maps corresponding to GC-Net$_{\text{det}}$ contain several artefacts with a very large disparity error that can be identified by their colouring in dark red and black and are visible in the examples of all four datasets, the results of the probabilistic adaptation contain clearly less such problematic regions. Additionally taking into account the corresponding region masks, it becomes evident that most of these artefacts are located in weakly textured regions of the

Table 6.3: **Quantitative comparison of the deterministic and probabilistic variant of GC-Net.** The probabilistic variant proposed in this work is compared against the deterministic baseline presented by Kendall et al. (2017b) based on the disparity error metrics described in Section 5.4.1.

	Pixel Error Rate [%]			MAE	RMSE
	$\tau = 1$	$\tau = 3$	$\tau = 5$	[px]	[px]
Sceneflow					
deterministic	10.4	5.5	4.3	1.8	10.5
probabilistic	12.1	5.6	3.7	0.9	3.7
KITTI					
deterministic	64.3	33.1	22.2	6.1	14.7
probabilistic	64.9	33.2	20.0	3.7	7.0
Middlebury					
deterministic	34.2	18.4	13.5	4.7	13.1
probabilistic	42.8	24.4	16.8	3.9	7.3
InStereo2K					
deterministic	65.9	44.4	36.3	17.5	34.1
probabilistic	71.7	51.0	41.0	10.0	15.9

(a) Trained on Sceneflow

	Pixel Error Rate [%]			MAE	RMSE
	$\tau = 1$	$\tau = 3$	$\tau = 5$	[px]	[px]
Sceneflow					
deterministic	24.0	11.7	9.6	3.9	13.3
probabilistic	23.9	15.0	12.5	4.6	14.2
KITTI					
deterministic	35.8	7.5	4.0	1.5	4.2
probabilistic	34.3	8.8	5.2	1.7	5.4
Middlebury					
deterministic	32.4	20.0	16.0	4.5	10.9
probabilistic	33.9	20.4	16.2	4.4	11.0
InStereo2K					
deterministic	17.9	8.9	7.1	3.7	8.0
probabilistic	19.1	10.4	7.8	2.5	5.7

(b) Fine-tuned on InStereo2K

reference image. Under the assumption that potentially matching pixels located in such a region are assigned similar cost values, the correct match can only be identified by chance. However, averaging the disparities assigned to these potentially matching pixels, the risk of choosing an incorrect pixel far away from the correct solution and thus making a gross mistake can be reduced. Consequently, the superior performance of GC-Net$_{\text{prob}}$ with respect to weakly textured image regions can mainly be explained by the Monte Carlo sampling approach applied to approximate the distribution of the disparity prediction.

To obtain a deeper insight into the differences between the deterministic and the probabilistic variant and to better understand which effects arise from the Monte Carlo sampling and which are due to the stochastic nature of the employed BNN approach, GC-Net$_{\text{prob}}$ is additionally examined using a single Monte Carlo sample instead of 50 as proposed in Section 5.4.5. As shown in Table 6.4, this probabilistic variant using a single Monte Carlo sample estimates disparities with similar pixel error rates as the deterministic baseline. The MAE and the RMSE, however, are slightly improved, which implies that the stochastic nature of this approach itself is already beneficial to reduce the number of large errors. In comparison, GC-Net$_{\text{prob}}$ using 50 Monte Carlo samples demonstrates a further clear improvement with respect to the MAE and the RMSE, while the pixel error rates deteriorate, as already described above. This behaviour is also visible in Figure 6.10, which shows the distribution of the estimated disparities with respect to the corresponding ground truth values. While an optimal dense stereo matching method would have an associated distribution that shows a thin diagonal line, the three evaluated variants show a relatively strong correlation in this sense but also a noticeable amount of larger deviations. On the one hand, GC-Net$_{\text{det}}$ tends to rather underestimate the true disparity, on the other hand, the probabilistic variant using a single Monte Carlo sample shows a more balanced distribution with a similar amount of deviations above and below the diagonal. More important is the fact that most deviations of the latter are clearly smaller than for

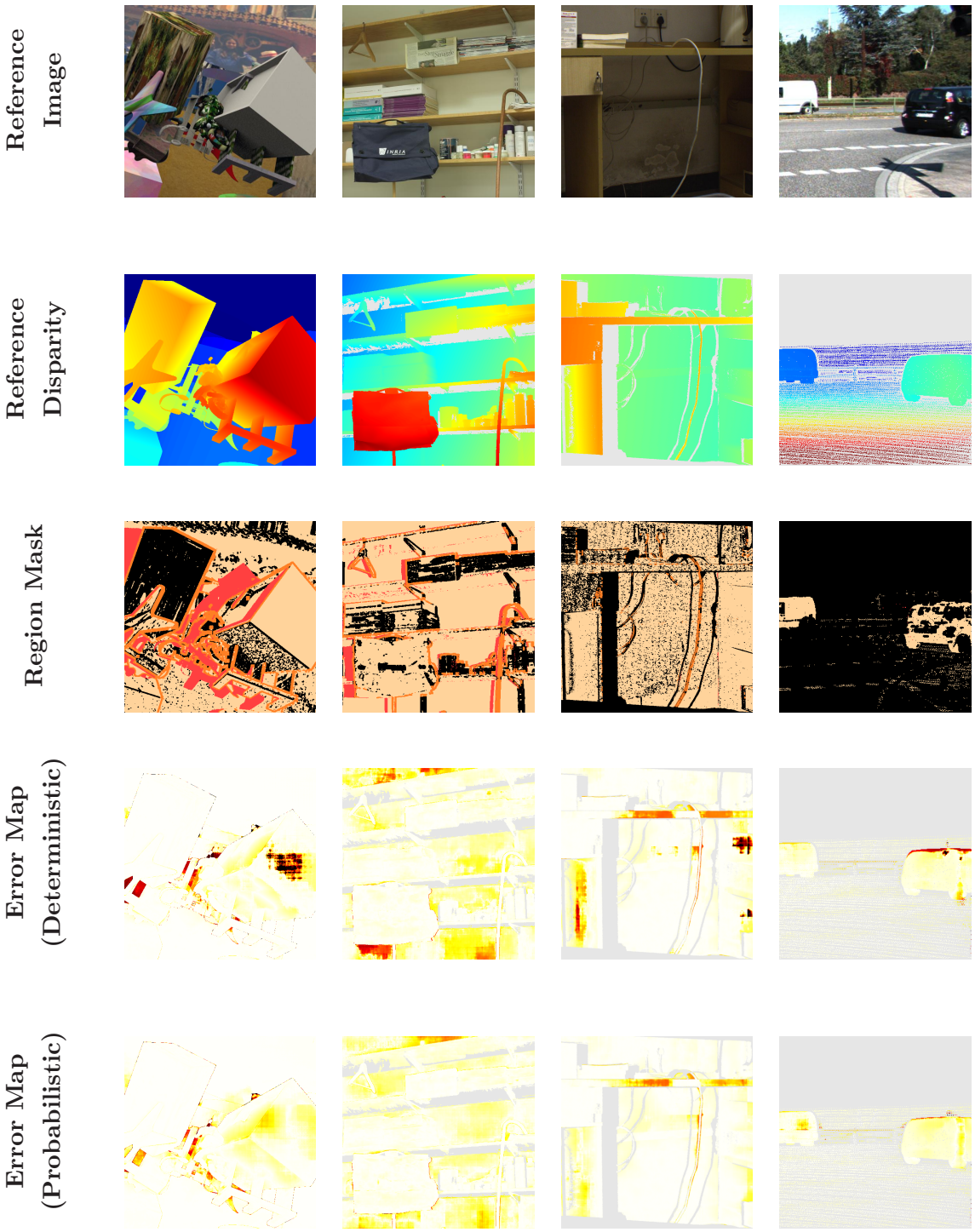


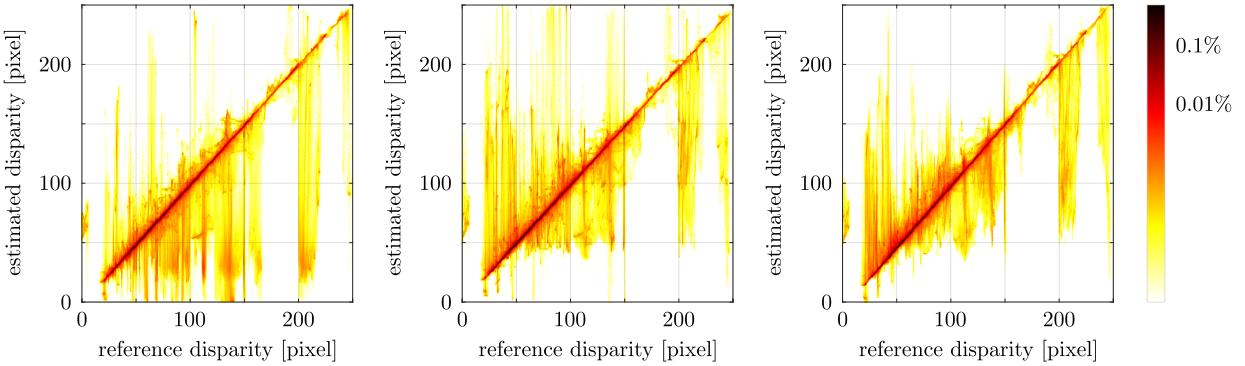
Figure 6.9: **Qualitative comparison of the disparity error.** From left to right, the figure shows an example for the Sceneflow, Middlebury, InStereo2K and KITTI dataset, respectively. The reference disparity maps shows small values in dark blue and large ones in red, while grey indicates pixels with unknown reference disparity. The region masks highlights regions that are especially challenging in the context of dense stereo matching, showing weakly textured areas in beige, occluded areas in red and pixels close to depth discontinuities in orange. The error maps show small values in white and large ones in dark red / black.

Table 6.4: **Evaluation of the effects arising from the probabilistic convolutional layers and the Monte Carlo sampling.** Three variants of GC-Net are compared, the deterministic baseline and the probabilistic variant presented in this work, using a single Monte Carlo sample or 50 samples (as proposed in Sec. 5.4.5), respectively. All variants are fine-tuned and tested on the InStereo2K dataset. Besides the disparity error, the training time in seconds per sample, the inference time in seconds per image and the memory footprint during training are listed for every variant.

	Pixel Error Rate [%]			MAE	RMSE	Training	Inference	Memory footprint
	$\tau = 1$	$\tau = 3$	$\tau = 5$	[px]	[px]	time [s]	time [s]	[GB]
deterministic	17.9	8.9	7.1	3.7	8.0	0.41	5	5.0
probabilistic (1 MC sample)	18.1	8.8	6.8	2.8	7.7	0.55	7	5.2
probabilistic (50 MC samples)	19.1	10.4	7.8	2.5	5.7	0.55	351	5.2

the deterministic variant. This applies even more for GC-Net$_{\text{prob}}$ using 50 Monte Carlo samples, for which the distribution is more centralised along the diagonal. Consequently, the conclusion drawn based on the numeric results that both, the stochastic nature as well as the sampling approach used in the probabilistic variant, contribute to improve the average disparity error, are also supported by the analysis of the correlation between estimated and reference disparity. Moreover, these results indicate that the increased number of small errors, as expressed by the higher pixel error rates for small thresholds, is purely caused by the strategy employed to aggregate the results of the individual Monte Carlo samples. To overcome this drawback, it may be beneficial to determine the mode of the posterior distribution as final disparity estimate instead of using the mean. However, because the approximation of the posterior is computed based on samples, the result is typically a discrete distribution which further depends on hyper-parameter settings, such as the number and width of bins used for the discretisation, which may lead to different modes for the same set of samples. Thus, further investigations are required in this context that need to be carried out in future work.

To close the comparison between the deterministic baseline and the probabilistic variant of GC-Net presented in this work, some practical aspects are examined. For this purpose, the training time per sample (an image extract of size 512×128 pixels as described in Sec. 5.3.3), the inference time per image (with respect to the images of the InStereo2K dataset as specified in Tab. 5.1) and the memory footprint during training are determined as the average values over all training or test samples of the InStereo2K dataset, respectively, and are listed in Table 6.4. To minimise the risk of effects that arise from anything else than the probabilistic adaptation proposed in this work, both variants, GC-Net$_{\text{det}}$ and GC-Net$_{\text{prob}}$, are implemented using the same framework (Tensorflow 2.3) and are trained and tested on the same hardware (AMD Ryzen 7 2700X 8×3.7 GHz, 32 GB memory, Nvidia Titan V). Analysing the training time and the memory footprint, it can be seen that the stochastic forward pass together with the Bayesian backpropagation employed to optimise the weights of the variational distribution of GC-Net$_{\text{prob}}$ are computationally more complex and require more memory than their deterministic counter-parts. This becomes particularly evident when keeping in mind that the number of feature channels and thus the size of the intermediate feature



(a) Deterministic (b) Probabilistic (1 sample) (c) Probabilistic (50 samples)

Figure 6.10: **Distribution of the estimated disparity with respect to the reference.** Three variants of GC-Net are compared, the deterministic baseline and the probabilistic variant presented in this work, using a single Monte Carlo sample and 50 samples (as proposed in Sec. 5.4.5), respectively. All variants are fine-tuned and tested on the InStereo2K dataset. The logarithmic colour scale denotes the percentage of pixels assigned the respective estimated and ground truth disparity, while a thin diagonal line would indicate an optimal result.

maps is reduced by 25% for the probabilistic variant compared to the baseline (see Sec. 4.3.1). In turn, this reduction helps to mitigate the difference between both variants, leading to an increase of training time of about 34% and a similar memory footprint. Note that the training time and the memory footprint are identical for the two probabilistic variants, because they only differ in the number of Monte Carlo samples used to approximate the posterior distribution, which only affects the behaviour at test time and has no influence on the training characteristics.

In contrast, the comparison of the inference time reveals significant differences and demonstrates a drawback of approaches that are based on Monte Carlo sampling: While the difference between the deterministic and the probabilistic variant using a single Monte Carlo sample is comparable to the one on the training time, the inference time increases linearly with the number of samples drawn. Thus, the probabilistic variant using 50 Monte Carlo samples requires about six minutes to compute a single disparity map, while this task is completed by the baseline in about five seconds. It may be argued that such a long inference time limits the practical relevance of the proposed approach, but it should be noted that several options exist to optimise the inference time: The computation of the individual samples can be fully parallelised or a more light-weight model could be used as basis (e.g. the adapted variant of GC-Net presented by Tulyakov et al. (2018)). Alternatively, the number of Monte Carlo samples drawn can be reduced if uncertainty estimates of lower accuracy are acceptable (cf. Fig. 5.3). Finally, first approaches are presented in the literature that allow to estimate epistemic uncertainty without relying on any kind of sampling, which clearly reduces the computational over-head (Postels et al., 2019).

6.3.2 On the Relevance of Aleatoric and Epistemic Uncertainty

In this second part of the experimental evaluation of the BNN presented in this thesis, the focus is put on the estimated uncertainty. For this purpose, four different variants of GC-Net are compared:

the deterministic baseline proposed by Kendall et al. (2017b), the probabilistic adaptation presented in Section 4.3 and both of these variants in combination with CVA-Net, as described in Section 4.4. Based on this comparison, the effects of exclusively estimating either aleatoric (deterministic GC-Net + CVA-Net) or epistemic uncertainty (probabilistic GC-Net) or predicting both of them jointly (probabilistic GC-Net + CVA-Net) are investigated. Besides the examination of the uncertainty, also the influence of combining GC-Net and CVA-Net on the disparity estimation capability is analysed. To allow a fair comparison, all four models are trained on the Sceneflow dataset first and are then fine-tuned on the InStereo2K dataset, as described in Sections 5.3.3 and 5.3.4 and as shown in Table 5.2b.

Analysing the correlation coefficients listed in Table 6.5, it can be seen that the variant that considers aleatoric and epistemic uncertainty jointly results in the highest correlation between the absolute disparity error and the estimated uncertainty given as standard deviation for all three datasets evaluated. While the exclusive consideration of epistemic uncertainty leads to slightly worse results, only taking into account aleatoric uncertainty reduces the correlation significantly. It is also noticeable that the correlation decreases, with an increase of the domain gap between training and test data. While the correlations are highest on the InStereo2K dataset which was also used for fine-tuning the network parameters, they are worse for the Middlebury dataset, which also shows indoor scenes as the images of the InStereo2K dataset but with different characteristics and captured using a different set-up (cf. Sec. 5.2), and are worst for the KITTI dataset, which shows outdoor scenes and thus has the largest domain gap to the training data. In addition, the variant that only estimates aleatoric uncertainty seems to be especially sensitive regarding these differences in the data processed. This effect can be explained by the fact that such a domain gap is mainly reflected by the uncertainty inherent in the model, because the definition of domain gap implies that the statistical properties of the data used to train the parameters of a model differs from the properties of the data used to test this model. In contrast, sources of uncertainty that are considered as being aleatoric are less affected, because concepts such as occlusion or texture remain unchanged for different datasets. Nevertheless, the ability of a model to estimate aleatoric uncertainty may very well be affected by a domain gap, leading to an indirect impact on the quality of the aleatoric uncertainty estimates. Consequently, because the uncertainty that is inherent in the model is neglected, the variant that considers aleatoric uncertainty only is less suitable to estimate uncertainty that arises from a domain gap in the data.

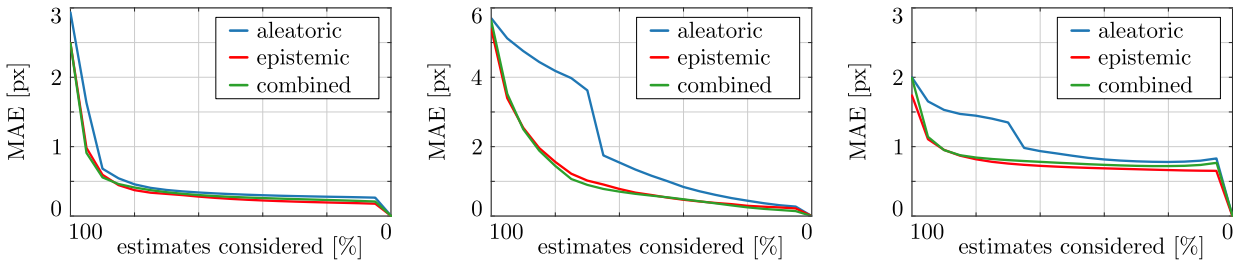
These observations are also supported by the sparsification plots shown in Figure 6.11. In these plots, the mean absolute error is shown with respect to the percentage of disparity estimates considered, which is reduced discarding pixels having assigned the largest uncertainty estimates first. While all three variants, estimating aleatoric or epistemic uncertainty only or considering both of them jointly, lead to similar curves for the InStereo2K dataset, significant differences can be seen for the Middlebury and the KITTI dataset. For these two datasets, the exclusive consideration of aleatoric uncertainty is not sufficient to infer the disparity error from the uncertainty, leading to a clearly higher MAE for the same density compared to the two other variants. This behaviour is also illustrated by the qualitative examples shown in Figures 6.12 and 6.13. For the example from the InStereo2K dataset, the uncertainty estimates of all three variants allow to identify the majority of erroneous disparity estimates, most of them being part of an artefact located at the

Table 6.5: **Comparison of different uncertainty models in the context of GC-Net.** The listed models are analysed with respect to the disparity error metrics described in Section 5.4.1, the average of the estimated disparity, whereas the combined standard deviation is computed based on Equation 4.21, and the Pearson correlation coefficient of the absolute disparity error and the combined estimated standard deviation (see Eq. 5.5). A hyphen indicates that a certain type of uncertainty is not estimated using the respective model.

	Pixel Error Rate [%]			MAE	RMSE	Standard deviation [px]			$r_{\Delta d,\sigma}$
	$\tau = 1$	$\tau = 3$	$\tau = 5$	[px]	[px]	alea.	epis.	comb.	
InStereo2K									
deterministic	17.9	8.9	7.1	3.7	8.0	-	-	-	-
deterministic + CVA-Net	18.1	9.0	7.3	2.9	6.8	1.3	-	1.3	0.57
probabilistic	19.1	10.4	7.8	2.5	5.7	-	2.4	2.4	0.69
probabilistic + CVA-Net	18.3	9.7	7.5	2.5	6.2	1.2	2.1	2.8	0.70
Middlebury									
deterministic	32.4	20.0	16.0	4.5	10.9	-	-	-	-
deterministic + CVA-Net	35.7	23.4	19.2	5.7	12.8	2.7	-	2.7	0.33
probabilistic	33.9	20.4	16.2	4.4	11.0	-	3.0	3.0	0.65
probabilistic + CVA-Net	35.8	23.7	19.4	5.6	12.1	1.8	2.8	4.0	0.70
KITTI									
deterministic	35.8	7.5	4.0	1.5	4.2	-	-	-	-
deterministic + CVA-Net	37.9	9.7	6.0	2.0	5.9	3.5	-	3.5	0.29
probabilistic	34.3	8.8	5.2	1.7	5.4	-	5.4	5.4	0.62
probabilistic + CVA-Net	36.2	9.7	6.1	2.0	6.1	2.2	5.3	6.7	0.66

left side of the image which is caused by the complete absence of texture in this region. With respect to the example from the KITTI dataset, however, only the variants that consider epistemic uncertainty are capable of predicting uncertainty estimates that show a strong relation to the actual disparity error. In contrast, the uncertainty map obtained with the variant that considers aleatoric uncertainty only contains higher uncertainties for more distant points in the scene, but does not provide particularly large uncertainty estimates for pixels with a large disparity error.

Focusing on the correlation coefficients in Table 6.5 that correspond to the variant that estimates aleatoric uncertainty only, it is noticeable that these values are clearly smaller than the ones obtained with the same aleatoric uncertainty model (the scene-aware model) using cost volumes originating from Census-based block matching or MC-CNN, as shown in Table 6.2. These differences may partially be caused by the domain gap between training and test data as discussed before. However, such a strong decrease of the correlation cannot be observed on the results presented in Section 6.2, although the impact of different datasets was also investigated there, training the approach on the KITTI dataset while using images from the Middlebury dataset for testing. Yet, a decrease of correlation can be observed with respect to the dense stereo matching method used to compute the cost volume which serves as basis for the estimation of aleatoric uncertainty using the approach presented in this thesis. More precisely, the more accurate the disparity maps obtained with a certain dense stereo matching method, the lower the achieved correlation between the estimated uncertainty and the disparity error. It can thus be assumed, that the quality of the disparity estimates has a direct influence on the ability to estimate the associated aleatoric uncertainty. This behaviour can be explained by the unbalanced distribution of the disparity error:

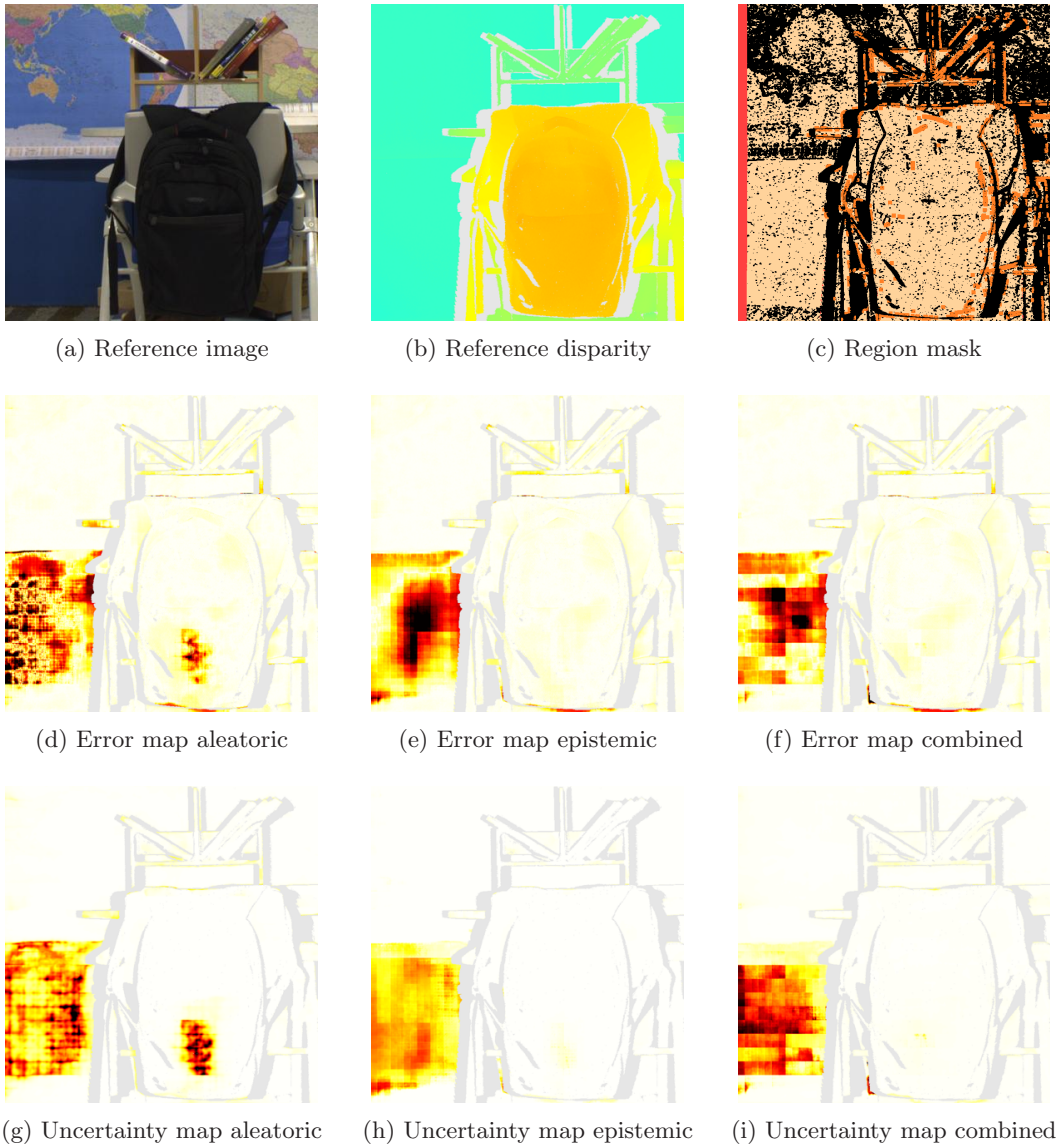


(a) InStereo2K (b) Middlebury (c) KITTI

Figure 6.11: **Sparsification plots with respect to the different types of uncertainties estimated.** The figures show the mean absolute disparity error considering all test images of the respective dataset on the y-axis and the percentage of considered disparity estimates on the x-axis. The percentage is reduced discarding pixels that have assigned the highest uncertainties first. The depicted curves correspond to the aleatoric, epistemic and combined uncertainty predicted with the deterministic GC-Net + CVA-Net, the probabilistic GC-Net and the probabilistic GC-Net + CVA-Net, respectively.

The more accurate a disparity map, the more pixels have assigned estimates with a disparity error close to zero. In turn, only a small percentage of pixels remain with a larger disparity error, which motivates the network to preferably learn the prediction of small uncertainties, because the chance for a correct guess is higher than for the prediction of a large uncertainty. Because this problem is similar to the one of imbalanced classes in the context of classification tasks, for the experiments conducted in the context of this thesis, it is addressed by weighting training samples in a manner inverse proportional to the frequency of their associated disparity error in the loss function (see Sec. 5.3.4). However, the effect that arises from such an imbalanced distribution of the disparity error is only mitigated but not resolved, as is clearly visible in Figure 6.14, in particular on the Middlebury dataset indicated by the horizontal line. Consequently, the issues of imbalanced training data in the context of regression tasks in general and the estimation of aleatoric uncertainty related to disparity maps of high accuracy in particular require further investigations. The correlation plots shown in Figure 6.14 as well as the associated correlation coefficients listed in Table 6.5 imply that the joint estimation of aleatoric and epistemic uncertainty leads to the best results and is thus the means of choice. Consequently, despite the remaining open challenges discussed, the consideration of aleatoric uncertainty is still useful, also in the context of the highly accurate disparity estimates obtained with GC-Net.

Analysing the mean standard deviations listed in Table 6.5 that correspond to the three different variants, several interesting effects can be observed. First of all, it can be seen that the estimated aleatoric and epistemic uncertainty is always larger if only one type of uncertainty is considered in the estimation. This indicates that while aleatoric and epistemic uncertainty can be clearly separated in theory, to some extent both approaches are able to also account for uncertainty from sources assigned to the respective other type of uncertainty. However, the standard deviation of their combination is always larger than the individual uncertainties, implying that both types of uncertainty contribute to an accurate quantification and that a model that takes into account only aleatoric or only epistemic uncertainty is not capable to reflect the error distribution to be expected properly. As discussed before, this observation is also supported by the respective correlation



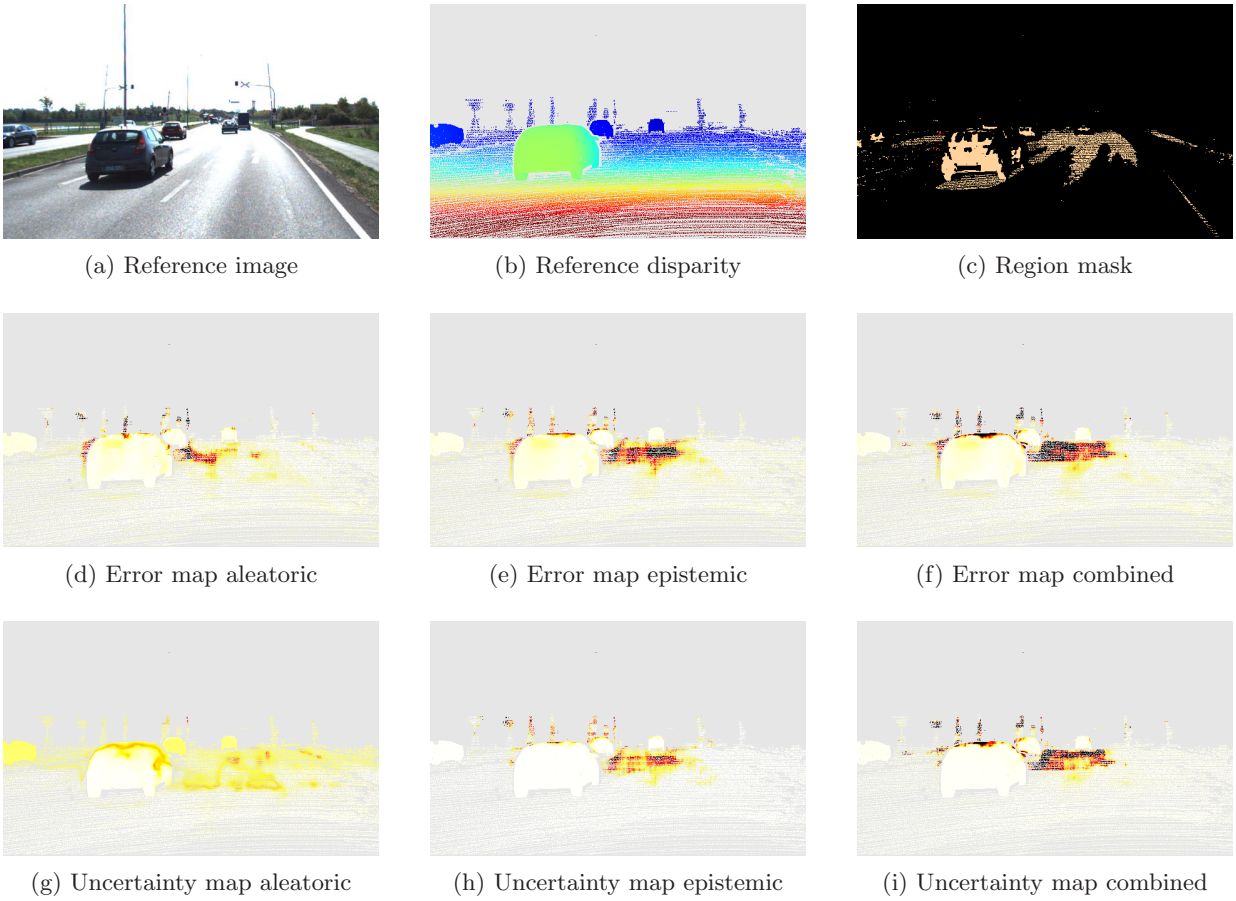
(a) Reference image (b) Reference disparity (c) Region mask

(d) Error map aleatoric (e) Error map epistemic (f) Error map combined

(g) Uncertainty map aleatoric (h) Uncertainty map epistemic (i) Uncertainty map combined

Figure 6.12: **Qualitative comparison of aleatoric and epistemic uncertainty and their combined estimation on an example of the InStereo2K dataset.** The figure shows the absolute disparity error estimated with the deterministic GC-Net + CVA-Net, the probabilistic GC-Net and the probabilistic GC-Net + CVA-Net in comparison to the associated uncertainty maps. In both the error and the uncertainty maps, small values are shown in white, large ones in dark red / black. Note that the values of the three uncertainty maps are scaled to the same interval to allow for an easier comparison. The reference disparity map shows large disparities in orange to red and small ones in turquoise to dark blue. The region mask highlights regions that are especially challenging in the context of dense stereo matching, showing weakly textured areas in beige, occluded areas in red and pixels close to depth discontinuities in orange.

coefficients. Moreover, it can be seen that the epistemic uncertainty dominates its aleatoric counterpart for all datasets, which is also clearly visible in the correlation plots shown in Figure 6.14. While a contribution of the aleatoric uncertainty is nevertheless observable when estimating both kinds of uncertainties jointly, the results might be further improvable if a weighting scheme is considered in the context of their aggregation. The results corresponding to the disparity error



(a) Reference image (b) Reference disparity (c) Region mask

(d) Error map aleatoric (e) Error map epistemic (f) Error map combined

(g) Uncertainty map aleatoric (h) Uncertainty map epistemic (i) Uncertainty map combined

Figure 6.13: **Qualitative comparison of aleatoric and epistemic uncertainty and their combined estimation on an example of the KITTI dataset.** The figure shows the absolute disparity error estimated with the deterministic GC-Net + CVA-Net, the probabilistic GC-Net and the probabilistic GC-Net + CVA-Net in comparison to the associated uncertainty maps. For an explanation of the colour coding, refer to Figure 6.12.

reveal that the combination with CVA-Net does not only allow to assess the aleatoric uncertainty, but also influences the disparity estimation itself. While for the deterministic variant of GC-Net this influence can mainly be seen on the improved MAE and RMSE values for the InStereo2K dataset, the combination with CVA-Net has a positive effect on the pixel error rates for the probabilistic variant of GC-Net (cf. Tab. 6.5). However, such an improvement can only be seen on the InStereo2K dataset that was used for fine-tuning the network parameters. In contrast, the disparity error is slightly increased for the Middlebury and the KITTI dataset. This indicates that the combination of GC-Net and CVA-Net leads to an over-fitting of the trainable parameters to the characteristics of the training data, which has a negative impact on the transferability of a trained model to other datasets.

Overall, the experimental results analysed in this section demonstrate the importance of estimating both aleatoric and epistemic uncertainty, in order to achieve an accurate and reliable estimation of the actual uncertainty associated to a depth estimate obtained via dense stereo matching. In practical terms, the large advantage of uncertainty estimation can be seen in the sparsification plots: Discarding only the 10% of pixels having assigned the highest uncertainty, the mean abso-

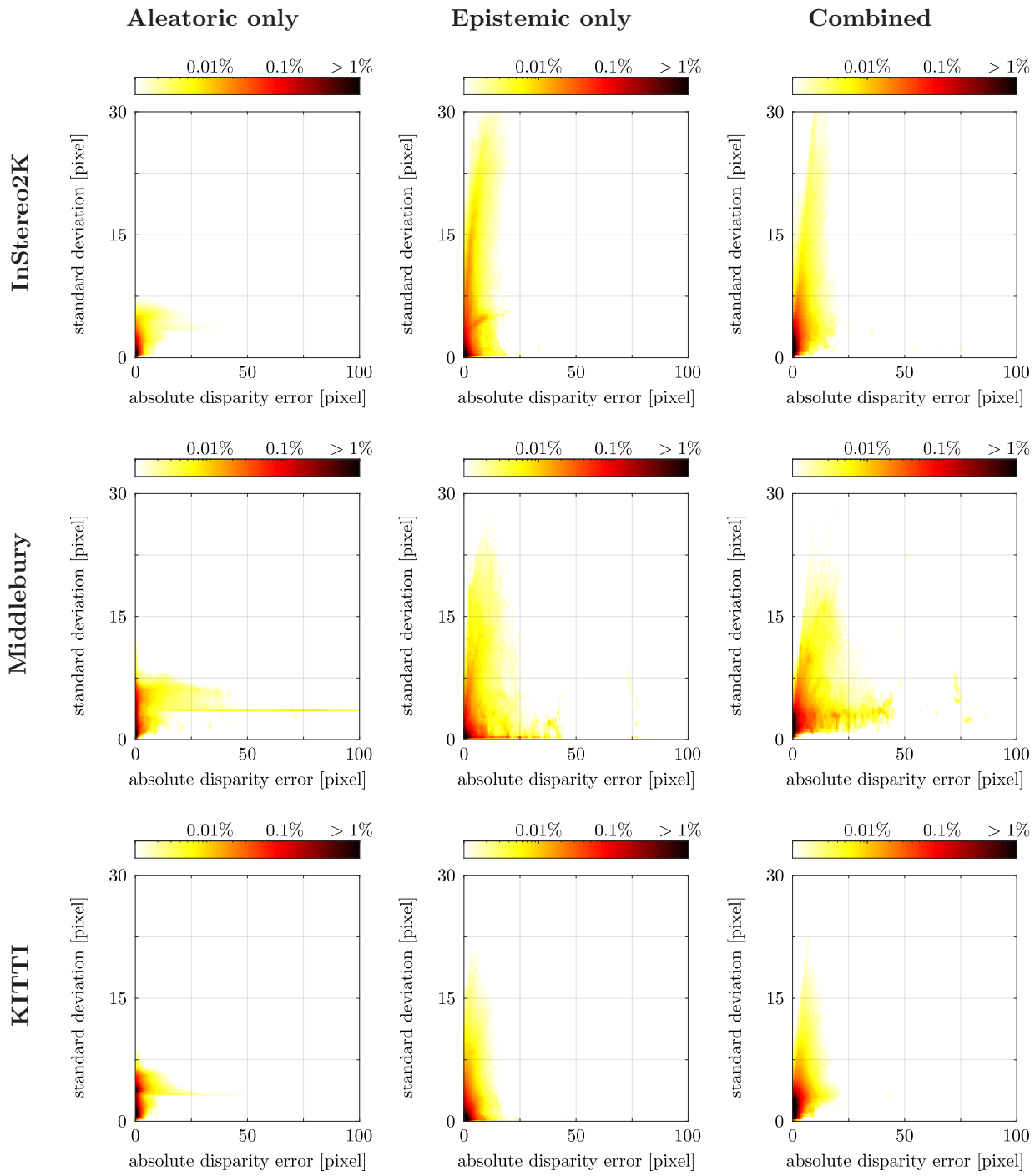


Figure 6.14: **Distributions of the different types of estimated uncertainties.** The figure shows the aleatoric, epistemic and combined uncertainty predicted with the deterministic GC-Net + CVA-Net, the probabilistic GC-Net and the probabilistic GC-Net + CVA-Net, respectively, with respect to the absolute disparity error using a logarithmic colour scale.

lute disparity error can be reduced by more than 50%, which is true for all datasets evaluated. This demonstrates that the approach for jointly estimating aleatoric and epistemic uncertainty presented in this work is capable of identifying the majority of erroneous disparity estimates and to assign an uncertainty with a magnitude that is related to the actual error magnitude as implied by the relatively high correlation coefficients achieved.

6.3.3 The Kullback-Leibler Divergence and the Mode Collapse Problem

In this section, an ablation study is presented, addressing the influence of the KL divergence on the results of the probabilistic variant of GC-Net if considered in the associated optimisation objective (cf. Eq. 4.19). As described in Section 2.3.2, the KL divergence works as a regularisation term, aiming to minimise the difference between the exact posterior distribution associated to the network parameters and the learned variational distribution used to approximate this posterior. However, as for most regularisation terms, the KL divergence is not necessary for a BNN to converge, which holds also true for the particular case of the probabilistic variant of GC-Net presented in this work. Why the integration of the KL divergence into the loss function is nevertheless a crucial part of the proposed approach is discussed in the following.

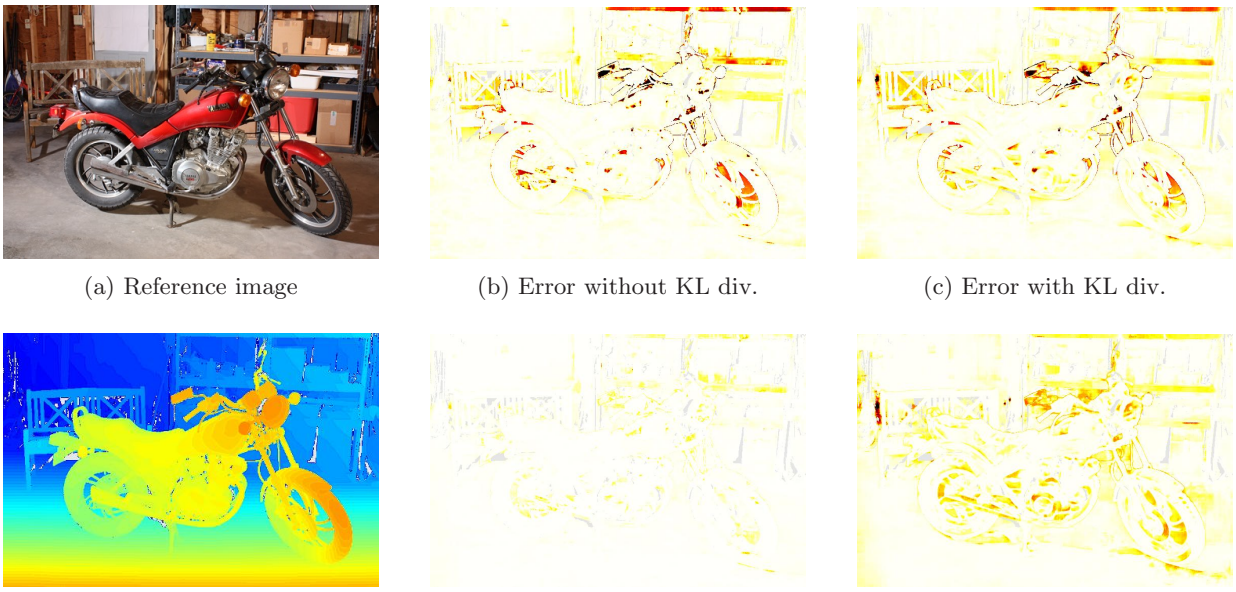
Analysing the quantitative comparison of the variants that consider or neglect the KL divergence in the loss function, respectively, shown in Table 6.6, it can be seen that the consideration of the KL divergence has a negative impact on the quality of the disparity estimates as indicated by the higher pixel error rates and the slightly increased mean absolute error. In contrast, the root mean squared error is equal for both variants, which in combination with the previous observations indicates that the variant neglecting the KL divergence suffers from a higher amount of large errors. From a general perspective, this negative influence on the quality can be explained by the fact that the consideration of the KL divergence requires to find a trade-off between the L1-norm used to minimise the disparity error and the deviation of the variational distribution from the exact posterior to obtain a minimal loss. More precisely, this behaviour may indicate that the choice of the prior distribution is not optimal, because the KL divergence is computed using the ELBO which in turn relies directly on the specified prior distribution. However, for a deeper insight into the effect of the prior distribution chosen, additional investigations are necessary that are to be carried out in future work.

In addition to the disparity error evaluated based on the different metrics described in Section 5.4.1, Table 6.6 further contains the average of the standard deviations learned as part of the variational distribution from which the network parameters of the probabilistic convolutional layers are randomly drawn. Comparing these average values belonging to the variants that consider or neglect the KL divergence, respectively, it is obvious that the value corresponding to the latter variant is significantly smaller and shows a clear deviation from the prior distribution which has a standard deviation of one. Consequently, the distribution corresponding to the variant that neglects the KL divergence is very narrow, in the literature sometimes also referred to as mode collapse. Sampling from such a narrow distribution, the network parameters obtained for different samples are rather similar which leads to only minor deviations in the disparity estimates computed based on these parameters. Thus, the epistemic uncertainty obtained with the model that neglects the KL divergence if often very low.

That such a mode collapse is an actual problem can be seen in Figure 6.15: While the epistemic uncertainty computed based on the variant that considers the KL divergence allows to identify the majority of erroneous disparity estimates, this is not the case for the variant that neglects the KL divergence. The latter one clearly under-estimates the disparity error, resulting in small epistemic

Table 6.6: **Quantitative results on the effect of the KL divergence in the optimisation objective.** The table shows the results of the probabilistic variant of GC-Net proposed, considering or neglecting the KL divergence in the loss function, evaluated on the Sceneflow dataset with respect to the disparity error metrics described in Section 5.4.1. In the last column, the average over the learned standard deviations σ_i describing the variational distribution (see Sec. 4.3.2) is listed.

	Pixel Error Rate [%]			MAE [px]	RMSE [px]	Avg. σ
	$\tau = 1$	$\tau = 3$	$\tau = 5$			
with KL divergence	12.1	5.6	3.7	0.9	3.7	0.73
without KL divergence	8.0	3.5	2.4	0.7	3.7	0.05



(a) Reference image (b) Error without KL div. (c) Error with KL div.

(d) Reference disparity (e) Uncertainty without KL div. (f) Uncertainty with KL div.

Figure 6.15: **Qualitative evaluation of the effect of the KL divergence on the estimated epistemic uncertainty.** The figure shows an example from the Middlebury dataset, comparing the epistemic uncertainty maps estimated based on the probabilistic variant of GC-Net presented in this work, while the KL divergence is either considered or neglected in the optimisation objective. In addition, the absolute differences between the estimated and the reference disparities are shown as error maps. While the reference disparity map shows small disparities in dark blue and large ones in orange, the error and uncertainty maps show small values in white and large ones in dark red / black. Analysing the uncertainty maps with respect to the corresponding disparity error maps, it can be seen that a majority of erroneous pixels can be identified based on the uncertainty map resulting from the variant that considers the KL divergence, while this is not the case for the variant that neglects the KL divergence.

uncertainty values. This finding is further supported by the calibration plot shown in Figure 6.16, which sets into relation the probability of a disparity estimate with the frequency of its actual occurrence. In this context, the probability is determined based on the estimated uncertainty, used to parameterise a Gaussian distribution over the disparity error, thus posing an expectation on the error distribution. Consequently, an optimal solution is achieved if the probability and frequency match, resulting in a diagonal in the calibration plot. Analysing this calibration plot, it can be seen that the variant that neglects the KL divergence leads to a frequency of disparity errors that is

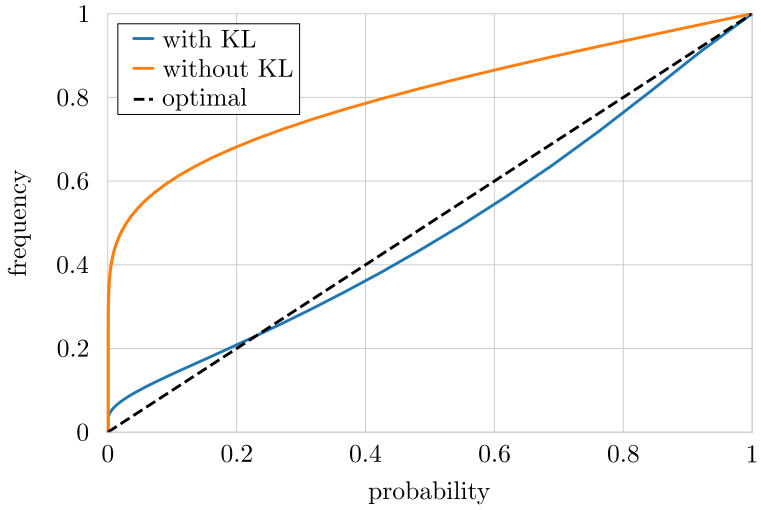


Figure 6.16: **Illustration of the uncertainty calibration.** The calibration plot shows the probability of the disparity error according to the predicted uncertainty on the x-axis in relation to the actual error distribution on the y-axis. Evaluated on the test samples of the Sceneflow dataset, the figure illustrates that the consideration of the KL divergence in the optimisation objective is crucial to obtain well-calibrated uncertainty estimates. On the other hand, if the KL divergence is neglected, the network tends to be overly optimistic with respect to the uncertainty, predicting values that are often significantly too small.

always significantly higher than the corresponding probability. Having in mind that the probability is defined based on a Gaussian, this behaviour shows that many more large disparity errors occur than implied by the uncertainty estimation. Thus, the variant that neglects the KL divergence tends to predict epistemic uncertainty values that are significantly too small. On the other hand, the variant that considers the KL divergence results in a graph that is very close to the diagonal, indicating that this model is well calibrated and predicts uncertainty values that match the actual disparity error in most cases.

6.4 Discussion

To close the chapter on the experimental results, in this section, the major findings of the investigations carried out are summarised, remaining open problems that became evident in the analysis of the results are discussed and the objectives of the experimental evaluation described in Section 5.1 are reviewed regarding their fulfilment. Moreover, the strengths and limitations identified on a theoretical basis in the methodology chapter are validated based on the practical evaluation.

Aleatoric Uncertainty Estimation

The first objective of the experiments was to evaluate the suitability of the proposed CVA-Net architecture to serve as functional model for the task of aleatoric uncertainty estimation. As demonstrated by the results of the experiments described in Section 6.1, this objective can be considered as fulfilled. The comparison against state-of-the-art methods from the literature has shown

that CVA-Net is able to achieve results of comparable quality and even surpasses these methods on some of the configurations evaluated. In this context, superior accuracy is particularly observable in noisy regions of a disparity map, caused, for example, by weak texture in the reference image or by occlusion. This behaviour further illustrates the advantages of processing cost volume-based information as realised by CVA-Net in comparison to relying on disparity map-based features: While several ambiguities that occur in the matching process due to such commonly challenging scenarios cannot be identified based on the information contained in the disparity map only, they may be obvious in the corresponding cost curves, e.g., in form of a wide and flat global minimum. Depth discontinuities, on the other hand, remain an open problem. CVA-Net as well as the other methods evaluated, often fail to estimate the uncertainty of pixels located close to such depth discontinuities correctly, whereas CVA-Net seems to be especially sensitive regarding this kind of situation, noticeable by the generally larger margin of incorrect estimates around depth discontinuities. However, the evaluation of the stochastic models presented for the task of aleatoric uncertainty estimation (see Sec. 6.2) has revealed that the use of CVA-Net in combination with probabilistic models allows to quantify the uncertainty reasonably well also for pixels located close to depth discontinuities. It can thus be concluded that the limitation of CVA-Net, handling depth discontinuities, is only partially caused by the architecture or the character of the features used, but also reinforced by the formulation of the uncertainty quantification task as confidence estimation via binary classification. Finally, the experimental results have shown that the concept of learning to estimate the uncertainty of a disparity assignment based on its cost curve generalises well over different datasets and can successfully be applied to different dense stereo matching methods.

The second objective of the experimental evaluation addressed the analysis of the described stochastic models for estimating the aleatoric uncertainty. As described on a theoretical basis in Section 4.2.2 and validated with experimental results in Section 6.2, a uni-modal distribution, such as a Laplacian, is only of limited suitability to represent the uncertainty associated to a pixel's depth estimate. While the usage of such a distribution is reasonable in the context of pixels for which no ambiguities exist regrading the identification of the correct correspondence, it does not fit to the actual error distributions in weakly textured or occluded regions of an image. The proposed combination with a uniform distribution, however, allows to overcome this limitation by explicitly addressing effects that cause a deviation from the assumption of a strictly uni-modal distribution. As demonstrated by the experimental results, this approach allows for a better quantification of the uncertainty, in such challenging regions as well as in total. In addition, two possibilities to combine the two types of probability distributions are proposed: The first variant is purely data driven, i.e., it learns the parameters characterising the two distributions as well as their weighting from training data. The second variant also learns the parameters of the distributions from training data, but ties the decision of which type of distribution to use for a particular pixel to scene characteristics. Comparing these two variants, the experimental results revealed that the combination of a data driven with a model-based approach as realised by the second variant leads to superior performance, observable for all evaluated configurations. Finally, it has to be noted that while the combination of a Laplace and a uniform distribution is sufficient to represent the uncertainty associated to the majority of pixels accurately, some cases remain for which this combination is not optimal, for example, pixels that are part of a repetitive pattern or that are located close to depth discontinuities.

While the latter case was also evaluated in the experiments conducted, showing results with good accuracy, it is to be expected that the usage of multi-modal distributions further improves the quality of the uncertainty estimates in such areas.

Epistemic and Joint Uncertainty Estimation

After the functional and the stochastic model developed for the purpose of aleatoric uncertainty estimation are examined in the first two sections of the experimental results, the third section focuses on the BNN proposed to estimate epistemic uncertainty and on the combination with CVA-Net to allow for a joint estimation of aleatoric and epistemic uncertainty. The experimental results presented in this section demonstrate the advantages of jointly estimating aleatoric and epistemic uncertainty in the context of dense stereo matching. As shown in Figure 6.11, the mean absolute disparity error can be halved for all datasets evaluated if only 10% of the pixels having assigned the highest uncertainties are discarded. This shows the ability of the proposed approach to identify the majority of erroneous disparity estimates correctly and to put them in a relative order that properly reflects the magnitude of their associated disparity error. In this context, the importance of the consideration of epistemic uncertainty is to be emphasised, in particular if a relevant domain gap exists between training and test data. While aleatoric uncertainty alone does not allow to infer the disparity error from the estimated uncertainty in such a scenario properly, epistemic uncertainty shows a strong correlation with the disparity error for the majority of pixels. However, also the estimation of aleatoric uncertainty contributes to the assessment of the overall uncertainty associated to a certain disparity estimate, as particularly shown by the superior correlation coefficients of the variant that estimates aleatoric and epistemic uncertainty jointly.

Despite these convincing results, the experimental evaluation also revealed some challenges that remain at least partially open and pose promising directions for future research. Comparing the results of the aleatoric uncertainty evaluation discussed in Section 6.2 and the results on the aleatoric uncertainty obtained in the context of the GC-Net-based experiments analysed in Section 6.3.2, it can be seen that the quality of the uncertainty estimates decreases with an increase in the accuracy of the disparity estimates originating from different dense stereo matching methods. It is to be assumed that this behaviour is at least partially caused by the imbalanced distribution of the disparity error in the training samples, leading to aleatoric uncertainty estimates that are systematically too small. In addition, the analysis of the influence of the consideration of the KL divergence in the loss function has shown that the joint estimation of disparity and epistemic uncertainty requires a trade-off that has a slightly negative impact on the accuracy of the disparity estimates (see Sec. 6.3.3). In turn, these results show that the presented BNN has the potential to further improve the disparity estimation capability. Since the KL divergence measures the deviation of the variational distribution from the exact posterior using the ELBO (cf. Sec. 2.3.2) which itself relies on the assumed prior for the variational distribution, it can be expected that a better choice for the prior positively affects the approximation of the posterior and thus also the overall results. In this context, especially the ability to consider correlations between individual parameters of a network are an exciting direction for future research.

Disparity Estimation

While the major objective of this work is to enable the estimation of uncertainty in the context of dense stereo matching, this should of course not be achieved at the expense of compromising the disparity estimation quality. Thus, it is desired that the methodology proposed allows to obtain disparity maps with a quality that is at least comparable to the one of the baseline. In general, it can be stated that this objective is achieved as demonstrated by the results discussed in Section 6.3.1. With respect to the mean disparity error and the number of large errors, the proposed probabilistic variant of GC-Net shows superior performance on several of the configurations evaluated. However, it must also be noted that the applied sampling strategy, or more precisely the aggregation scheme used to combine the disparity estimates of the individual Monte Carlo samples, has a slightly negative impact on the quality. These findings are consistent over the different datasets evaluated and can be observed independent of whether the deterministic and probabilistic variants are trained on synthetic data only or additionally fine-tuned on real data.

Practical Limitations

As demonstrated by the results of the experimental evaluation, the presented methodology achieves state-of-the-art accuracy with respect to the estimated uncertainty and fulfils the objectives stated in the context of this thesis. However, CVA-Net, the architecture presented for the purpose of aleatoric uncertainty estimation, as well as the probabilistic adaptation of GC-Net are mainly based on 3D convolutions which leads to a relatively high computational burden and a large memory footprint. In addition, the BNN-based approach followed to enable the estimation of epistemic uncertainty, further increases the training and inference time compared to the deterministic baseline (cf. Tab. 6.4). Together with the already higher number of epochs needed for the training procedure to converge, the training time is enlarged significantly. While the proposed reduction of the model size using a lower number of filter channels per layer helps to mitigate this effect, it is still clearly noticeable. While it can be argued that the training time is only of minor importance for the practical relevance of a method, also the inference time is increased, especially due to the sampling approach employed to approximate the exact posterior distribution. However, as described in Section 6.3.1, the implementation of the method presented was not explicitly optimised with the goal of minimising runtime or memory consumption and several promising options exist to overcome these practical limitations in future work.

7 Conclusions and Outlook

Addressing the task of uncertainty estimation in the context of dense stereo matching, a holistic approach is presented in this thesis that allows depth to be jointly estimated along with its associated uncertainty based on a stereo image pair. In this last chapter, conclusions are drawn regarding the presented approach as well as the experimental results. After a brief summary of strengths and limitations of the individual components and the approach as a whole, remaining open issues are identified, providing an outlook on promising future research topics.

Inspired by the convincing results of learned uncertainty measures on the one hand and uncertainty estimation based on features from cost curves on the other hand as described in the literature, in this thesis, a novel CNN architecture called CVA-Net is developed to learn the estimation of aleatoric uncertainty in the context of dense stereo matching based on 3D cost volumes. It is argued that such cost volumes contain additional information compared to disparity maps, which allows to estimate the associated uncertainty more accurately, in particular for pixels without a distinct correspondence, for example, due to weak texture or occlusions in the reference image. In addition to this architecture, also an approach to use mixture distributions to describe the aleatoric uncertainty is presented. With this approach, the limitations of confidence estimation, being unable to quantify the uncertainty in pixels or metric units, and uni-modal probability distributions, assuming a unique optimum in the matching process, can be overcome. More precisely, it is proposed to combine a Laplace and a uniform distribution, following two different concepts, either learning the combination of these two distributions based on training data or deciding per pixel which distribution to be used based on scene characteristics. Based on the experimental results, it can be concluded that the latter concept, combining a data-driven strategy to learn the prediction of the distribution parameters with a model-based approach used to tie the type of distribution to certain assumptions on the scene properties, is the means of choice. In general, the usage of such a mixture distribution clearly improves the quality of the uncertainty estimates, both in regions that violate the assumption implied by uni-modal distributions and overall. While both the CNN architecture as well as the uncertainty models presented for the task of aleatoric uncertainty estimation demonstrate convincing results and surpass state-of-the-art methods from the literature in the majority of configurations evaluated, some limitations remain and pose promising starting points for potential future works: Depth discontinuities are known to be particularly challenging in the context of dense stereo matching and have also demonstrated to be characterised as such for the task of aleatoric uncertainty estimation. To overcome this problem, it may be beneficial to additionally learn features on the reference image, allowing CVA-Net to better localise depth discontinuities and to identify foreground fattening effects based on image gradients. The additional consideration of such 2D features from the reference image or the disparity map could also allow to enlarge the receptive field without increasing the computational burden as much as purely

operating on three dimensional data. From the perspective of the stochastic model, multi-modal distributions, such as Gaussian mixture models, may allow to quantify the uncertainty in these regions more accurately and further allow for a higher flexibility, which may also be beneficial for other challenging scenarios, such as repetitive patterns. Another possible direction for further research on the topic of aleatoric uncertainty estimation could address the propagation of uncertainty through a neural network. This would especially be interesting if prior information on the uncertainty is available, for example, obtained during the camera calibration, and could thus be considered in the uncertainty estimation process.

To additionally allow for the estimation of epistemic uncertainty, the use of a BNN is proposed in this thesis. For this purpose, a conventional deterministic CNN architecture from the literature, which has already proven to be well-suited for the task of dense stereo matching, is transformed into a Bayesian representation, using probabilistic convolutional layers which are trained via variational inference. At test time, the characteristics of the distribution associated to the disparity estimates is approximated via Monte Carlo sampling. In contrast to other approaches for the estimation of epistemic uncertainty, such as Monte Carlo dropout or ensemble learning, the usage of a BNN provides a higher flexibility with respect to the definition of prior distributions and the consideration of correlations between the parameters of a neural network. However, this theoretical freedom is limited due to practical aspects such as the number of trainable parameters, which is increased significantly if covariances are considered in the estimation, or the complexity of the resulting variational distribution which is to be optimised during the training process. Due to these limitations, a naive mean-field approximation with a Gaussian prior and a diagonal covariance matrix is used as stochastic model for the proposed BNN architecture in this work. While the definition of a suitable prior for the variational distribution is not a trivial task, it plays a decisive role for the accuracy of the depth and uncertainty estimates obtained with such a BNN, as implied by the ablation study on the relevance of the KL divergence. Consequently, both, further investigations on the definition of the prior and the consideration of correlations, for example, extending the mean-field approximation to a general formulation, are exciting directions for future research. However, the experimental results have nevertheless clearly shown the advantages of estimating the epistemic uncertainty, especially if a relevant domain gap exists between training and test data. While the exclusive consideration of aleatoric uncertainty is sufficient to detect erroneous disparity estimates in the absence of such a domain gap, it does not allow to capture the model uncertainty which typically dominates the uncertainty arising from the data given a strong difference in the characteristics of training and test data. Overall, the joint estimation of both aleatoric and epistemic uncertainty has demonstrated the best results and is thus the means of choice. However, drawbacks of the proposed approach are the enlarged memory footprint, the higher number of parameters to be trained as well as the significantly increased training and inference time. To address these issues, in this work it is proposed to only transform a subset of all convolutional layers into a probabilistic representation and to reduce the number of filter channels in all convolutional layers. Both adaptations together lead to a memory footprint and a number of trainable parameters of the proposed BNN that are comparable to the deterministic baseline, without having a negative impact on the accuracy of the disparity or uncertainty estimates. Yet, the problem of the increased training and inference time remains, and particularly the latter is highly relevant if the proposed approach is to

be incorporated into any practical application as long as hardware limitations prevent a parallel computation of the individual Monte Carlo samples. Based on the observation that a large share of the inference time is caused by the sampling procedure used, the investigation of approaches that allow to obtain a good approximation of the exact posterior distribution with a reduced number of Monte Carlo samples or without relying on any sampling approach at all are a promising direction for future work.

Besides the specific suggestions for future work discussed so far, also more general possibilities exist to extend the concept of uncertainty estimation presented in this thesis. While the presented approach is limited to processing stereo image pairs consisting of two images, also an extension to multi-view stereo is conceivable. Especially in the context of an actual 3D reconstruction (rather than the 2.5D representation of the scene geometry provided by a disparity or depth map), such an extension might be of interest to improve the fusion of depth information defined with respect to different reference images or to estimate the uncertainty associated to a surface reconstruction. In addition, also temporal aspects may be taken into account, extending the approach presented to process pairs of image sequences instead of single image pairs, which would build the basis for the estimation of uncertainty associated to optical flow or scene flow. Finally, addressing the estimation of semantic information and dense stereo matching together has shown to improve the results of both of these tasks, which makes it also interesting to incorporate semantic information into the presented uncertainty estimation approach. The combination of uncertainties related to continuous variables such as depth and categorical ones such as semantic labels is a challenging task, indeed, and worth investigating.

In conclusion, it can be stated that with the methodology presented in this thesis, the objectives set out at the beginning are achieved. The developed concept for estimating the uncertainty in the context of dense stereo matching can be seen as a further step towards the comprehensive understanding of this topic and can serve as a strong foundation for promising future research addressing the discussed possibilities.

Bibliography

Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S. and Zhang, X., 2020. InStereo2K: A Large Real Dataset for Stereo Matching in Indoor Scenes. *Science China Information Sciences*. 60

Batsos, K. and Mordohai, P., 2018. RecResNet: A Recurrent Residual CNN Architecture for Disparity Map Enhancement. In: *Proceedings of the International Conference on 3D Vision*, pp. 238–247. 23

Batsos, K., Cai, C. and Mordohai, P., 2018. CBMV: A Coalesced Bidirectional Matching Volume for Disparity Estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2060–2069. 20, 27

Birchfield, S. and Tomasi, C., 1998. A Pixel Dissimilarity Measure that is Insensitive to Image Sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(4), pp. 401–406. 19

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer, New York. 12, 18

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D., 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* 112(518), pp. 859–877. 17

Bleyer, M., Rhemann, C. and Rother, C., 2011. PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In: *Proceedings of the British Machine Vision Conference*, pp. 14.1–14.11. 25

Blundell, C., Cornebise, J., Kavukcuoglu, K. and Wierstra, D., 2015. Weight Uncertainty in Neural Networks. In: *Proceedings of the International Conference on Machine Learning*, pp. 1613–1622. 18, 32, 34

Boykov, Y., Veksler, O. and Zabih, R., 1999. Fast Approximate Energy Minimization via Graph Cuts. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 377–384. 8

Breiman, L., 1996. Bagging Predictors. *Machine Learning* 24(2), pp. 123–140. 31

Brosse, N., Riquelme, C., Martin, A., Gelly, S. and Moulines, É., 2020. On Last-Layer Algorithms for Classification: Decoupling Representation from Uncertainty Estimation. *arXiv preprint arXiv:2001.08049*. 32, 34, 49

Bulatov, D., Wernerus, P. and Heipke, C., 2011. Multi-view Dense Matching supported by Triangular Meshes. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(6), pp. 907–918. 10, 21

Chang, J.-R. and Chen, Y.-S., 2018. Pyramid Stereo Matching Network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418. 24

Chen, L., Rottensteiner, F. and Heipke, C., 2020. Deep Learning Based Feature Matching and Its Application in Image Orientation. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2020, pp. 25–33. 14

Chen, Z., Sun, X., Wang, L., Yu, Y. and Huang, C., 2015. A Deep Visual Correspondence Embedding Model for Stereo Matching Costs. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 972–980. 20, 21

Cheng, X., Wang, P. and Yang, R., 2019. Learning Depth with Convolutional Spatial Propagation Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42(10), pp. 2361–2379. 24

Der Kiureghian, A. and Ditlevsen, O., 2008. Aleatory or Epistemic? Does it matter? *Structural Safety* 31, pp. 105–112. 10, 11

Dietterich, T. G., 2002. Ensemble Learning. In: M. A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, second edn, MIT Press, Cambridge. 17

Ditlevsen, O. and Madsen, H. O., 1996. *Structural Reliability Methods*. Wiley New York. 11

Duggal, S., Wang, S., Ma, W.-C., Hu, R. and Urtasun, R., 2019. Deep-Pruner: Learning Efficient Stereo Matching via Differentiable Patchmatch. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4384–4393. 25

Felzenszwalb, P. F. and Huttenlocher, D. P., 2006. Efficient Belief Propagation for Early Vision. *International Journal of Computer Vision* 70(1), pp. 41–54. 21

Förstner, W., 1982. On the Geometric Precision of Digital Correlation. *International Archives of Photogrammetry and Remote Sensing* 24(3), pp. 176–189. 7

Fu, Z., Ardabilian, M. and Stern, G., 2019. Stereo Matching Confidence Learning based on Multi-Modal Convolution Neural Networks. *Representations, Analysis and Recognition of Shape and Motion from Imaging Data* pp. 69–81. 27, 28, 43, 71, 72

Gal, Y. and Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In: *Proceedings of the International Conference on Machine Learning*, pp. 1050–1059. 31, 34

Gast, J. and Roth, S., 2018. Lightweight Probabilistic Deep Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3369–3378. 30

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. 59

Geyer, C., 2011. Introduction to Markov Chain Monte Carlo. In: S. Brooks, A. Gelman, G. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, pp. 3–48. 17

Ghiasi, G., Lin, T.-Y. and Le, Q. V., 2018. DropBlock: A Regularization Method for Convolutional Networks. *Advances in Neural Information Processing Systems* 31, pp. 10750–10760. 32

Gidaris, S. and Komodakis, N., 2017. Detect, Replace, Refine: Deep Structured Prediction for Pixel Wise Labeling. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5248–5257. 23

Girshick, R., 2015. Fast R-CNN. *Proceedings of the IEEE International Conference on Computer Vision* pp. 1440–1448. 14, 64

Glorot, X. and Bengio, Y., 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 249–256. 12, 62

Glorot, X., Bordes, A. and Bengio, Y., 2011. Deep Sparse Rectifier Neural Networks. In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 315–323. 12, 16

Goodfellow, I., Bengio, Y. and Courville, A., 2016. *Deep Learning*. MIT Press. 12

Graves, A., 2011. Practical Variational Inference for Neural Networks. *Advances in Neural Information Processing Systems* 24, pp. 2348–2356. 17, 18, 32, 51

Gul, M. S. K., Bätz, M. and Keinert, J., 2019. Pixel-Wise Confidences for Stereo Disparities Using Recurrent Neural Networks. In: *Proceedings of the British Machine Vision Conference*, pp. 23.1–23.13. 28

Guney, F. and Geiger, A., 2015. Displets: Resolving Stereo Ambiguities using Object Knowledge. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4165–4175. 10, 21

Guo, X., Yang, K., Yang, W., Wang, X. and Li, H., 2019. Group-Wise Correlation Stereo Network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3273–3282. 25

Hacking, I., 1975. *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge University Press. 2, 10, 11

Haeusler, R., Nair, R. and Kondermann, D., 2013. Ensemble Learning for Confidence Measures in Stereo Vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 305–312. 27

Hamzah, R. A., Ibrahim, H. and Hassan, A. H. A., 2017. Stereo Matching Algorithm based on per Pixel Difference Adjustment, Iterative Guided Filter and Graph Segmentation. *Journal of Visual Communication and Image Representation* 42, pp. 145–160. 21

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. 14

Heinrich, K. and Mehltretter, M., 2021. Learning Multi-Modal Features for Dense Matching-Based Confidence Estimation. *ISPRS Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B2-2021, pp. 91–99. 54

Helton, J. C., 1994. Treatment of Uncertainty in Performance Assessments for Complex Systems. *Risk Analysis* 14(4), pp. 483–511. 10

Henrion, M. and Fischhoff, B., 1986. Assessing Uncertainty in Physical Constants. *American Journal of Physics* 54(9), pp. 791–798. 10

Heo, Y. S., Lee, K. M. and Lee, S. U., 2011. Robust Stereo Matching using Adaptive Normalized Cross-Correlation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(4), pp. 807–822. 19

Hernández-Lobato, J. M. and Adams, R., 2015. Probabilistic Backpropagation for Scalable Learning Of Bayesian Neural Networks. In: *Proceedings of the International Conference on Machine Learning*, pp. 1861–1869. 34

Hirschmuller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), pp. 328–341. 10, 21, 23, 38, 62, 71, 72

Hirschmuller, H. and Scharstein, D., 2009. Evaluation of Stereo Matching Costs on Images with Radiometric Differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(9), pp. 1582–1599. 20

Hoffman, M. D., Blei, D. M., Wang, C. and Paisley, J., 2013. Stochastic Variational Inference. *Journal of Machine Learning Research* 14(5), pp. 1303–1347. 18, 32, 51

Höllmann, M., Mehltretter, M. and Heipke, C., 2020. Geometry-Based Regularisation for Dense Image Matching via Uncertainty-Driven Depth Propagation. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2020, pp. 151–159. 10

Hong, L. and Chen, G., 2004. Segment-based Stereo Matching using Graph Cuts. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 74–81. 10, 21

Hu, X. and Mordohai, P., 2012. A Quantitative Evaluation of Confidence Measures for Stereo Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(11), pp. 2121–2133. 20, 26, 29, 67

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E. and Weinberger, K. Q., 2017. Snapshot Ensembles: Train 1, Get m for Free. In: *Proceedings of the International Conference on Learning Representations*. 31

Huber, P. J., 1981. *Robust Statistics*. John Wiley & Sons, Inc., New York. 45

Ilg, E., Saikia, T., Keuper, M. and Brox, T., 2018. Occlusions, Motion and Depth Boundaries With a Generic Network for Disparity, Optical Flow or Scene Flow Estimation. In: *Proceedings of the European Conference on Computer Vision*, pp. 614–630. 25

Intille, S. S. and Bobick, A. F., 1994. Disparity-Space Images and Large Occlusion Stereo. In: *Proceedings of the European Conference on Computer Vision*, Springer, pp. 179–186. 7

Ioffe, S. and Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In: *Proceedings of the International Conference on Machine Learning*, pp. 448–456. 16

Jiang, H., Sun, D., Jampani, V., Lv, Z., Learned-Miller, E. and Kautz, J., 2019. Sense: A Shared Encoder Network for Scene-Flow Estimation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3195–3204. 25

Jie, Z., Wang, P., Ling, Y., Zhao, B., Wei, Y., Feng, J. and Liu, W., 2018. Left-Right Comparative Recurrent Model for Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3838–3846. 22

Jospin, L. V., Buntine, W., Boussaid, F., Laga, H. and Bennamoun, M., 2020. Hands-on Bayesian Neural Networks - a Tutorial for Deep Learning Users. *arXiv preprint arXiv:2007.06823*. 17, 18, 31

Kang, J., Chen, L., Deng, F. and Heipke, C., 2019. Context Pyramidal Network for Stereo Matching Regularized by Disparity Gradients. *ISPRS Journal of Photogrammetry and Remote Sensing* 157, pp. 201–215. 24, 25

Kendall, A. and Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems* 30, pp. 5574–5584. 10, 11, 29, 30, 31, 34, 43, 44

Kendall, A., Badrinarayanan, V. and Cipolla, R., 2017a. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In: *Proceedings of the British Machine Vision Conference*, pp. 57.1–57.12. 31, 34

Kendall, A. G., 2017. Geometry and Uncertainty in Deep Learning for Computer Vision. PhD thesis, University of Cambridge, Department of Engineering. 28, 30, 34

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A. and Bry, A., 2017b. End-to-End Learning of Geometry and Context for Deep Stereo Regression. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 66–75. 14, 15, 16, 24, 38, 43, 49, 65, 83, 84, 88

Khamis, S., Fanello, S., Rhemann, C., Kowdle, A., Valentin, J. and Izadi, S., 2018. Stereonet: Guided Hierarchical Refinement for Real-Time Edge-Aware Depth Prediction. In: *Proceedings of the European Conference on Computer Vision*, pp. 573–590. 25

Kim, S., Kim, S., Min, D. and Sohn, K., 2019a. LAF-Net: Locally Adaptive Fusion Networks for Stereo Confidence Estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 205–214. 28

Kim, S., Min, D., Ham, B., Kim, S. and Sohn, K., 2017. Deep Stereo Confidence Prediction for Depth Estimation. In: *Proceedings of the IEEE International Conference on Image Processing*, pp. 992–996. 28, 33

Kim, S., Min, D., Kim, S. and Sohn, K., 2019b. Unified Confidence Estimation Networks for Robust Stereo Matching. *IEEE Transactions on Image Processing* 28(3), pp. 1299–1313. 28, 29, 33, 62

Kim, S., Min, D., Kim, S. and Sohn, K., 2020. Adversarial Confidence Estimation Networks for Robust Stereo Matching. *IEEE Transactions on Intelligent Transportation Systems* pp. 1–15. 28

Kingma, D. P. and Ba, J., 2015. Adam: A Method for Stochastic Optimization. *Proceedings of the International Conference on Learning Representations*. 13, 63

Kingma, D. P., Salimans, T. and Welling, M., 2015. Variational Dropout and the Local Reparameterization Trick. *Advances in Neural Information Processing Systems* 28, pp. 2575–2583. 18, 32, 34, 51

Knöbelreiter, P. and Pock, T., 2019. Learned Collaborative Stereo Refinement. In: *Proceedings of the German Conference on Pattern Recognition*, Springer, Cham, pp. 3–17. 23

Knöbelreiter, P., Reinbacher, C., Shekhovtsov, A. and Pock, T., 2017. End-To-End Training of Hybrid CNN-CRF Models for Stereo. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2339–2348. 24

Kullback, S. and Leibler, R. A., 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1), pp. 79–86. 18

Kuschk, G., 2019. Efficient Large-Scale Stereo Reconstruction using Variational Methods. PhD thesis, Technical University of Munich, Department of Informatics. 21

Lakshminarayanan, B., Pritzel, A. and Blundell, C., 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. *Advances in Neural Information Processing Systems* 30, pp. 6402–6413. 31, 34

Leal-Taixé, L., Canton-Ferrer, C. and Schindler, K., 2016. Learning by Tracking: Siamese CNN for Robust Target Association. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 33–40. 14

Li, A., Chen, D., Liu, Y. and Yuan, Z., 2016a. Coordinating Multiple Disparity Proposals for Stereo Computation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4022–4030. 21

Li, L., Zhang, S., Yu, X. and Zhang, L., 2016b. PMSC: PatchMatch-based Superpixel Cut for Accurate Stereo Matching. *IEEE Transactions on Circuits and Systems for Video Technology* 28(3), pp. 679–692. 21

Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Zhou, L. and Zhang, J., 2018. Learning for Disparity Estimation through Feature Constancy. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2811–2820. 24, 25

Liu, J., Paisley, J., Kioumourtzoglou, M.-A. and Coull, B., 2019. Accurate Uncertainty Estimation and Decomposition in Ensemble Learning. *Advances in Neural Information Processing Systems* 32, pp. 8952–8963. 10

Long, J., Shelhamer, E. and Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440. 14

Luan, X., Yu, F., Zhou, H., Li, X., Song, D. and Wu, B., 2012. Illumination-robust Area-based Stereo Matching with Improved Census Transform. In: *International Conference on Measurement, Information and Control*, pp. 194–197. 20

Luo, W., Schwing, A. G. and Urtasun, R., 2016. Efficient Deep Learning for Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5695–5703. 20

MacKay, D. J. C., 1992. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation* 4(3), pp. 448–472. 32

Marmanis, D., Wegner, J. D., Galliani, S., Schindler, K., Datcu, M. and Stilla, U., 2016. Semantic Segmentation of Aerial Images with an Ensemble of CNNs. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* III-3, pp. 473–480. 14

Marr, D. and Poggio, T., 1979. A Computational Theory of Human Stereo Vision. *Proceedings of the Royal Society of London. Series B. Biological Sciences* 204(1156), pp. 301–328. 9

Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A. and Brox, T., 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4040–4048. 24, 26, 43, 59

Mehltretter, M., 2020. Uncertainty Estimation for End-To-End Learned Dense Stereo Matching via Probabilistic Deep Learning. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2020, pp. 161–169. 16, 38, 52, 70

Mehltretter, M. and Heipke, C., 2019. CNN-based Cost Volume Analysis as Confidence Measure for Dense Matching. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 2070–2079. 39, 40

Mehltretter, M. and Heipke, C., 2021. Aleatoric Uncertainty Estimation for Dense Stereo Matching via CNN-based Cost Volume Analysis. *ISPRS Journal of Photogrammetry and Remote Sensing* 171, pp. 63–75. 7, 38, 68

Mehltretter, M., Kleinschmidt, S. P., Wagner, B. and Heipke, C., 2018. Multimodal Dense Stereo Matching. In: *Proceedings of the German Conference on Pattern Recognition*, Springer, Cham, pp. 407–421. 20

Menze, M. and Geiger, A., 2015. Object Scene Flow for Autonomous Vehicles. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3061–3070. 43, 59

Mouats, T., Aouf, N. and Richardson, M. A., 2015. A Novel Image Representation via Local Frequency Analysis for Illumination Invariant Stereo Matching. *IEEE Transactions on Image Processing* 24(9), pp. 2685–2700. 9

Moukari, M., Simon, L., Picard, S. and Jurie, F., 2019. n-MeRCI: A new Metric to Evaluate the Correlation Between Predictive Uncertainty and True Error. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5250–5255. 31, 34

Neal, R. M., 1995. Bayesian Learning for Neural Networks. PhD thesis, University of Toronto. 32

Nguyen, U. and Heipke, C., 2020. 3D Pedestrian Tracking using Local Structure Constraints. *ISPRS Journal of Photogrammetry and Remote Sensing* 166, pp. 347–358. 14

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B. and Snoek, J., 2019. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. *Advances in Neural Information Processing Systems* 32, pp. 13991–14002. 31

Pang, J., Sun, W., Ren, J. S., Yang, C. and Yan, Q., 2017. Cascade Residual Learning: A Two-Stage Convolutional Neural Network for Stereo Matching. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 887–895. 24, 25

Park, H. and Lee, K. M., 2017. Look Wider to Match Image Patches With Convolutional Neural Networks. *IEEE Signal Processing Letters* 24(12), pp. 1788–1792. 21

Park, M.-G. and Yoon, K.-J., 2015. Leveraging Stereo Matching with Learning-based Confidence Measures. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 101–109. 22, 27

Pizzoli, M., Forster, C. and Scaramuzza, D., 2014. REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2609–2616. 30, 45

Poggi, M. and Mattoccia, S., 2016a. Deep Stereo Fusion: Combining Multiple Disparity Hypotheses with Deep-Learning. In: *Proceedings of the International Conference on 3D Vision*, pp. 138–147. 27, 29

Poggi, M. and Mattoccia, S., 2016b. Learning a General-purpose Confidence Measure based on O(1) Features and a smarter Aggregation Strategy for Semi Global Matching. In: *Proceedings of the International Conference on 3D Vision*, pp. 509–518. 27

Poggi, M. and Mattoccia, S., 2016c. Learning from Scratch a Confidence Measure. In: *Proceedings of the British Machine Vision Conference*, BMVA Press, pp. 46.1–46.13. 27, 28, 71, 72

Poggi, M. and Mattoccia, S., 2017. Learning to predict Stereo Reliability enforcing Local Consistency of Confidence Maps. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2452–2461. 27

Poggi, M., Kim, S., Tosi, F., Kim, S., Aleotti, F., Min, D., Sohn, K. and Mattoccia, S., 2021a. On the Confidence of Stereo Matching in a Deep-Learning Era: A Quantitative Evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 26, 28

Poggi, M., Pallotti, D., Tosi, F. and Mattoccia, S., 2019. Guided Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 979–988. 24, 26

Poggi, M., Tosi, F. and Mattoccia, S., 2017. Even More Confident Predictions with Deep Machine-Learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 76–84. 27

Poggi, M., Tosi, F., Batsos, K., Mordohai, P. and Mattoccia, S., 2021b. On the Synergies Between Machine Learning and Stereo: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24

Postels, J., Ferroni, F., Coskun, H., Navab, N. and Tombari, F., 2019. Sampling-free Epistemic Uncertainty Estimation Using Approximated Variance Propagation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2931–2940. 87

Riesch, H., 2012. Levels of Uncertainty. In: S. Roeser, R. Hillerbrand, P. Sandin and M. Peterson (eds), *Handbook of Risk Theory. Epistemology, Decision Theory, Ethics, and Social Implications of Risk.*, Vol. 1, Springer, Dordrecht, chapter 4, pp. 88–110. 10, 11

Ronneberger, O., Fischer, P. and Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham, pp. 234–241. 14

Rosasco, L., Vito, E. D., Caponnetto, A., Piana, M. and Verri, A., 2004. Are Loss Functions All the Same? *Neural Computation* 16(5), pp. 1063–1076. 13

Scharstein, D. and Szeliski, R., 2002. A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision* 47(1-3), pp. 7–42. 7, 19, 33, 37, 47, 69

Scharstein, D., Hirschmüller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X. and Westling, P., 2014. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. In: *Proceedings of the German Conference on Pattern Recognition*, Springer, Cham, pp. 31–42. 60

Schönberger, J. L., Sinha, S. N. and Pollefeys, M., 2018. Learning to Fuse Proposals from Multiple Scanline Optimizations in Semi-Global Matching. In: *Proceedings of the European Conference on Computer Vision*, pp. 739–755. 22, 76

Schuster, R., Wasenmuller, O., Unger, C. and Stricker, D., 2019. SDC - Stacked Dilated Convolution: A Unified Descriptor Network for Dense Matching Tasks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2556–2565. 21

Seki, A. and Pollefeys, M., 2016. Patch Based Confidence Prediction for Dense Disparity Map. In: *Proceedings of the British Machine Vision Conference*, BMVA Press, pp. 23.1–23.13. 22, 27

Seki, A. and Pollefeys, M., 2017. SGM-Nets: Semi-Global Matching With Neural Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 231–240. 22

Shaked, A. and Wolf, L., 2017. Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4641–4650. 28

Shimizu, M. and Okutomi, M., 2005. Sub-Pixel Estimation Error Cancellation on Area-Based Matching. *International Journal of Computer Vision* 63(3), pp. 207–224. 23

Song, X., Zhao, X., Hu, H. and Fang, L., 2018. EdgeStereo: A Context Integrated Residual Pyramid Network for Stereo Matching. In: *Proceedings of the Asian Conference on Computer Vision*, Springer, pp. 20–35. 25

Spyropoulos, A. and Mordohai, P., 2015. Ensemble Classifier for Combining Stereo Matching Algorithms. In: *Proceedings of the International Conference on 3D Vision*, pp. 73–81. 27

Spyropoulos, A., Komodakis, N. and Mordohai, P., 2014. Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1621–1628. 22, 27

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* 15(1), pp. 1929–1958. 31, 63

Stentoumis, C., Grammatikopoulos, L., Kalisperakis, I. and Karras, G., 2014. On Accurate Dense Stereo-Matching using a Local Adaptive Multi-Cost Approach. *ISPRS Journal of Photogrammetry and Remote Sensing* 91, pp. 29–49. 20, 23

Stucker, C. and Schindler, K., 2020. ResDepth: Learned Residual Stereo Reconstruction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 184–193. 23

Sullivan, T. J., 2015. *Introduction to Uncertainty Quantification*. Springer, Cham. 11

Sun, L., Chen, K., Song, M., Tao, D., Chen, G. and Chen, C., 2017. Robust, Efficient Depth Reconstruction with Hierarchical Confidence-Based Matching. *IEEE Transactions on Image Processing* 26(7), pp. 3331–3343. 27

Taniai, T., Matsushita, Y. and Naemura, T., 2014. Graph Cut based Continuous Stereo Matching using Locally Shared Labels. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1613–1620. 21

Tappen, M. F. and Freeman, W. T., 2003. Comparison of Graph Cuts with Belief Propagation for Stereo, Using Identical MRF Parameters. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 900–906. 8

Tieleman, T. and Hinton, G., 2012. Lecture 6.5 - RMSprop: Divide the Gradient by a Running Average of its Recent Magnitude. *COURSERA: Neural Networks for Machine Learning*. 13, 65

Tonioni, A., Poggi, M., Mattoccia, S. and Di Stefano, L., 2017. Unsupervised Adaptation for Deep Stereo. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1605–1613. 1, 26

Tonioni, A., Tosi, F., Poggi, M., Mattoccia, S. and Stefano, L. D., 2019. Real-Time Self-Adaptive Deep Stereo. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 195–204. 25

Tosi, F., Poggi, M. and Mattoccia, S., 2019. Leveraging Confident Points for Accurate Depth Refinement on Embedded Systems. *Proceedings of the IEEE Embedded Vision Workshop*. 23

Tosi, F., Poggi, M., Benincasa, A. and Mattoccia, S., 2018. Beyond Local Reasoning for Stereo Confidence Estimation with Deep Learning. In: *Proceedings of the European Conference on Computer Vision*, pp. 319–334. 27, 29, 43, 54, 71, 72

Tulyakov, S., Ivanov, A. and Fleuret, F., 2018. Practical Deep Stereo (PDS): Toward Applications-friendly Deep Stereo Matching. *Advances in Neural Information Processing Systems* 31, pp. 5871–5881. 25, 87

Van Asselt, M. B. A. and Rotmans, J., 2002. Uncertainty in Integrated Assessment Modelling. *Climatic Change* 54(1-2), pp. 75–105. 10, 11

Veld, R. O. H., Jaschke, T., Bätz, M., Palmieri, L. and Keinert, J., 2018. A Novel Confidence Measure for Disparity Maps by Pixel-Wise Cost Function Analysis. In: *Proceedings of the International Conference on Image Processing*, pp. 644–648. 26

Vogiatzis, G. and Hernández, C., 2011. Video-Based, Real-Time Multi-View Stereo. *Image and Vision Computing* 29(7), pp. 434–441. 30, 45

Wei, Y. and Quan, L., 2004. Region-Based Progressive Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 106–113. 21

Wen, Y., Vicol, P., Ba, J., Tran, D. and Grosse, R., 2018. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. In: *Proceedings of the International Conference on Learning Representations*. 32, 51

Wu, Z., Wu, X., Zhang, X., Wang, S. and Ju, L., 2019. Semantic Stereo Matching With Pyramid Cost Volumes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7484–7493. 25

Yamaguchi, K., McAllester, D. and Urtasun, R., 2014. Efficient Joint Segmentation, Occlusion Labeling, Stereo and Flow Estimation. In: *Proceedings of the European Conference on Computer Vision*, Springer, pp. 756–771. 21

Yang, G., Zhao, H., Shi, J., Deng, Z. and Jia, J., 2018. SegStereo: Exploiting Semantic Information for Disparity Estimation. In: *Proceedings of the European Conference on Computer Vision*, pp. 636–651. 25

Yang, Y., Yuille, A. and Lu, J., 1993. Local, Global, and Multilevel Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 274–279. 7

Yu, L., Wang, Y., Wu, Y. and Jia, Y., 2018. Deep Stereo Matching with Explicit Cost Aggregation Sub-Architecture. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7517–7524. 24

Zabih, R. and Woodfill, J., 1994. Non-Parametric Local Transforms for Computing Visual Correspondence. In: *Proceedings of the European Conference on Computer Vision*, Springer, Berlin, Heidelberg, pp. 151–158. 7, 9, 19, 38, 62, 71, 72, 77, 78

Zbontar, J. and LeCun, Y., 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *Journal of Machine Learning Research* 17(1), pp. 2287–2318. 20, 24, 38, 62, 71, 72, 77, 78

Zendel, O., Murschitz, M., Humenberger, M. and Herzner, W., 2017. How Good Is My Test Data? Introducing Safety Analysis for Computer Vision. *International Journal of Computer Vision* 125(1-3), pp. 95–109. 2, 26

Zeng, J., Lesnikowski, A. and Alvarez, J. M., 2018. The Relevance of Bayesian Layer Positioning to Model Uncertainty in Deep Bayesian Active Learning. In: *Proceedings of the Conference on Neural Information Processing Systems Workshop on Bayesian Deep Learning*. 32, 34, 49

Zhan, W., Ou, X., Yang, Y. and Chen, L., 2019. DSNet: Joint Learning For Scene Segmentation and Disparity Estimation. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2946–2952. 25

Zhang, C., Li, Z., Cai, R., Chao, H. and Rui, Y., 2014. As-Rigid-As-Possible Stereo under Second Order Smoothness Priors. In: *Proceedings of the European Conference on Computer Vision*, Springer, pp. 112–126. 21

Zhang, F. and Wah, B. W., 2017. Fundamental Principles on Learning New Features for Effective Dense Matching. *IEEE Transactions on Image Processing* 27(2), pp. 822–836. 20

Zhang, F., Prisacariu, V., Yang, R. and Torr, P. H. S., 2019. GA-Net: Guided Aggregation Net for End-to-End Stereo Matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 185–194. 24

Zhang, F., Qi, X., Yang, R., Prisacariu, V., Wah, B. and Torr, P., 2020. Domain-Invariant Stereo Matching Networks. *Proceedings of the Europe Conference on Computer Vision* pp. 420–439. 26

Zhang, K., Lu, J. and Lafruit, G., 2009. Cross-based Local Stereo Matching using Orthogonal Integral Images. *IEEE Transactions on Circuits and Systems for Video Technology* 19(7), pp. 1073–1079. 20

Zhang, S., Xie, W., Zhang, G., Bao, H. and Kaess, M., 2017. Robust Stereo Matching with Surface Normal Prediction. In: *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 2540–2547. 23

Zhong, Z. and Mehltretter, M., 2021. Mixed Probability Models for Aleatoric Uncertainty Estimation in the Context of Dense Stereo Matching. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* V-2-2021, pp. 17–26. 38

Zhou, Y. T. and Chellappa, R., 1988. Computation of Optical Flow using a Neural Network. In: *Proceedings of the IEEE International Conference on Neural Networks*, pp. 71–78. 14

Acknowledgements

As a dissertation is not just the achievement of an individual, but also the result of good advice, intensive discussions and substantial support, in the following, I would like to thank a number of people who accompanied me through my doctorate and contributed to the success of this work. In addition, I would like to thank the German Research Foundation (DFG), which supported me as part of the Research Training Group i.c.sens [GRK2159], as well as the MOBILISE initiative of the Leibniz University Hannover and TU Braunschweig for funding my research.

First of all, I would like to express my deep gratitude to my supervisor Prof. Dr.-Ing. habil. Christian Heipke. I thank him for the very pleasant collaboration, for always taking time, not only for questions with scientific content, but also for all the other challenges one encounters during a doctorate, and of course for the countless pieces of good advice he gave me along the way. I feel more than fortunate that he gave me the chance to come to IPI and to learn from him so much. Moreover, I would like to thank Prof. Dr. Konrad Schindler and Dr.-Ing. habil. Hamza Alkhatib, who acted as co-reviewers for this dissertation, for reviewing it and for their valuable comments and suggestions.

Many thanks also to my former office companions and friends Nicolas Garcia Fernandez, Max Coenen, Uyen Nguyen, Hani Dbouk, Gregor Blott, Philipp Trusheim and Sara El Amrani as well as my colleagues from IPI and i.c.sens for the countless discussions about science as well as life in general and for making it always a pleasure to come to work. Moreover, I would like to thank Sebastian Kleinschmidt and Raphael Voges as well as Prof. Dr. Jan Dirk Wegner for the exciting discussions and the collaborations we had during my doctorate.

Finally, I could not have completed my dissertation without the support of my family. Many thanks to my parents Lutz and Petra as well as my sister Marie for always supporting me and keeping me motivated to ask questions and to stay curious, and thus paving my way into science. To close, a particularly big thank you to my wife Romina for her unconditional love, for always believing in me and for taking care of me throughout this whole journey.