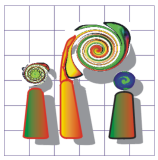


Gottfried Wilhelm Leibniz Universität Hannover
Institute of Photogrammetry and GeoInformation

MASTER THESIS

Mixed Probability Models for Aleatoric Uncertainty Estimation in the Context of Dense Stereo Matching

Zeyun Zhong
10018162

Supervisor:
M.Sc. Max Mehltretter

Examiner:
Prof. Dr.-Ing. habil. Christian Heipke
M.Sc. Max Mehltretter

Hannover, March 2021

Statement

I declare that this thesis is the result of independent research conducted by me under the guidance of my supervisor. It does not contain the results of any other scientific research that has been published or written by any other individuals or groups, except for those already cited in the thesis. Furthermore, I state that this work in the same or a similar form has not been submitted to an examination authority.

Signature

Place, Date

Abstract

The ability to identify erroneous depth estimates is of fundamental interest. Information regarding the aleatoric uncertainty of depth estimates can be, for example, used to support the process of depth reconstruction itself. Consequently, various methods for the estimation of aleatoric uncertainty in the context of dense stereo matching have been presented in recent years, with deep learning-based approaches being particularly popular. Among these deep learning-based methods, probabilistic strategies are increasingly attracting interest, because the estimated uncertainty can be quantified in pixels or in metric units due to the consideration of real error distributions. However, existing probabilistic methods usually assume a unimodal distribution to describe the error distribution while simply neglecting cases in real-world scenarios that could violate this assumption. To overcome this limitation, we propose three novel mixed probability models consisting of Laplacian and Uniform distributions as well as Gaussian mixture models for the task of aleatoric uncertainty estimation. In this way, we explicitly address commonly challenging regions in the context of dense stereo matching and outlier measurements, respectively. To allow a fair comparison, we adapt a common neural network architecture to investigate the effects of the different uncertainty models. In an extensive evaluation using two datasets and two common dense stereo matching methods, the proposed methods demonstrate state-of-the-art accuracy.

Keywords Uncertainty Quantification, 3D Reconstruction, Deep Learning, Mixture Model

Contents

1 Introduction **1**

2 Theoretical Background **3**

2.1 Basic Definitions . . . 3

2.2 Cost Computation . . . 4

2.2.1 Census Transform . . . 5

2.2.2 MC-CNN fast . . . 5

2.3 CVA-Net . . . 6

2.4 Probability Distributions . . . 7

2.4.1 Gaussian Distribution . . . 7

2.4.2 Laplacian Distribution . . . 8

2.4.3 Uniform Distribution . . . 10

2.4.4 Gaussian Mixture Distribution . . . 10

3 Related Works **12**

3.1 Uncertainty Estimation . . . 12

3.2 Assessment of Predictive Uncertainty . . . 13

4 Methodology **15**

4.1 Input and Problem Statement . . . 15

4.2 Basic Architecture . . . 15

4.3 Probabilistic Uncertainty Models . . . 17

4.3.1 Laplacian Model . . . 17

4.3.2 Geometry-aware Model . . . 18

4.3.3 GMM Model . . . 20

4.3.4 Laplacian Uniform Mixture Model . . . 23

5 Experimental Setup **24**

5.1 Datasets . . . 24

5.2 Binary Region Masks . . . 24

5.3 Training Procedure . . . 25

5.4 Metrics 25
5.4.1 AUC 26
5.4.2 Correlation Coefficient 27
5.4.3 n-MeRCI 27
5.5 Evaluation Procedure 28

6 Results **30**
6.1 AUC and Correlation Coefficient 30
6.1.1 Comparison of the Uncertainty Models 30
6.1.2 Investigation of the GMM Uncertainty Model 37
6.2 n-MeRCI 39

7 Conclusion **43**

Bibliography **44**

1 Introduction

Depth estimation from images, i.e., reconstructing the per-pixel distance between a scene and a camera, is a classical task in photogrammetry as well as in computer vision. It has many applications in practice, including fields of autonomous vehicles and UAVs (Achtelik et al., 2009). It also serves as a foundation to support other photogrammetric computer vision problems, such as 3D reconstruction (Newcombe et al., 2011) and object detection (Girshick, 2015).

As a central step of depth estimation in any photogrammetric 3D reconstruction, the core task of dense stereo matching is the determination of pixel correspondences for all pixels in an image pair. It is challenging to achieve this objective robustly in real-world scenarios due to a variety of challenges, such as occlusions, thin structures and large weakly textured areas, e.g. in the sky or caused by over-exposure. A typical behaviour for pixel in such regions is the absence of an identifiable global minimum in the cost curve, making the determination of the correct correspondence unreliable. It is therefore crucial to be able to assess how trustworthy this information is.

To address the task of uncertainty quantification in the context of dense stereo matching, in this work, we focus on the estimation of aleatoric uncertainty. From the perspective of dense stereo matching, aleatoric uncertainty accounts for effects such as sensor noise, occlusion, depth discontinuities and matching ambiguities caused by texture-less areas or repetitive patterns (Mehltretter and Heipke, 2021). In the literature, a variety of different deep learning-based approaches to estimate aleatoric uncertainty have been proposed in recent years demonstrating convincing results (Poggi and Mattoccia, 2016b; Mehltretter and Heipke, 2021). Among them, two types of strategies are especially popular: confidence-based and probabilistic-based.

Confidence-based methods predict a score between zero and one for each pixel, which indicates the trust that is put in a pixel's depth estimate and can be learned as a binary class probability of a depth estimate being correct or incorrect (Hu and Mordohai, 2012). In contrast, probabilistic methods assume a certain probability distribution for aleatoric uncertainty which is optimised during training maximising the likelihood (Kendall and Gal, 2017). While this approach requires more detailed knowledge on the real error distribution, contrary to the concept of confidence-based methods, it allows to additionally quantify the uncertainty in pixels or in metric units.

These probabilistic methods usually assume a unimodal distribution, considering the error distribution as a Gaussian (Kendall and Gal, 2017) or a Laplacian distribution (Mehltretter and Heipke,

2021). However, this is not always the case in the context of dense stereo matching and does especially not account for outlier measurements or commonly challenging regions, such as texture-less regions, occlusions and depth discontinuities.

To overcome these limitations, we propose three novel mixed probability models for aleatoric uncertainty estimation, using different combinations of a Laplacian and a Uniform distributions as well as Gaussian mixture models under varying assumptions. We evaluate our proposed methods together with the state-of-the-art probabilistic aleatoric uncertainty model, i.e., Laplacian model, on two different datasets, containing outdoor and indoor scenes, respectively. For a fair comparison, the *Cost Volume Analysis Network* (CVA-Net) (Mehltretter and Heipke, 2021) is adapted to investigate the differences and effects of these three uncertainty models. Thus, the main contributions of this work are:

- A geometry-aware probabilistic aleatoric uncertainty model that explicitly models regions that are challenging in the context of dense stereo matching.
- A GMM probabilistic aleatoric uncertainty model which explicitly addresses the multimodality in the context of dense stereo matching.
- A mixture probabilistic aleatoric uncertainty model explicitly considering outlier measurements.
- An adaption of the CVA-Net architecture for the proposed uncertainty models. In addition, the network architecture is optimised to accelerate the training procedure.
- An investigation of the uncertainty evaluation metric n-MeRCI in context of this work.

In the remainder of this thesis, we first describe the theoretical background in Chapter 2, including the basic definitions related to this work, two cost computation methods, the CVA-Net architecture and probability distributions, before the related work is reviewed in Chapter 3. Next, we introduce our proposed probabilistic uncertainty models in Chapter 4. The experimental setup is presented in Chapter 5, while results and limitations of our approach are discussed in Chapter 6. Finally, we conclude this work and give an outlook on future work in Chapter 7.

2 Theoretical Background

In this chapter, the theoretical background is described, in which the basic definitions supporting this work are introduced in Section 2.1. In the subsequent Section 2.2, two algorithms are reviewed for the first stage of many stereo algorithms: the matching cost computation. Based on the results of the matching cost computation, the CVA-Net is utilised to estimate aleatoric uncertainty of disparity estimates. Its architecture is briefly outlined in Section 2.3. In the last Section 2.4, the probability distributions relevant for this work are discussed.

2.1 Basic Definitions

Given two epipolar rectified images, we wish to compute the disparity d for each pixel in the left image in order to get the depth information z using the following relation:

$$z = \frac{fB}{d}, \tag{2.1}$$

where f is the focal length of the camera and B is the distance between the camera centers. Disparity d refers to the difference in horizontal location of a pixel at position (x, y) in the left and the corresponding pixel at position $(x - d, y)$ in the right image. An example of disparity map is illustrated in Figure 2.1c.

The first stage of many stereo algorithms to compute disparity map is the matching cost computation (Scharstein and Szeliski, 2002), in which the matching cost, i.e. the dissimilarity between a pixel of interest and a potential correspondence, for example, the squared difference of intensity values (SD), is computed for each pixel of the left image and for each disparity value within a specific range. This results in a 3D cost volume of size $w \times h \times |\delta|$ (c.f. Fig. 2.2), where the width w and height h correspond to the width and height of images and $|\delta|$ denotes the length of considered disparity range.

The cost curve of a pixel in the left image, i.e. the matching costs of potential correspondences in the right image along the epipolar line and within the specified disparity range $|\delta|$, contains a lot of information, such as position and number of local optima and the curvature etc., which can be used to determine the pixel correspondence and also to assess the reliability of the disparity

(a) Left image (b) Right image (c) Disparity map (d) Uncertainty map

Figure 2.1: Illustration of a stereo image pair and its corresponding disparity and uncertainty maps. Note that objects closer to the camera have larger disparities than objects farther away. The third image presents the disparity map with warmer colors representing larger values of disparity (and smaller values of depth). The uncertainty map encodes a high value in black and a small one in white.

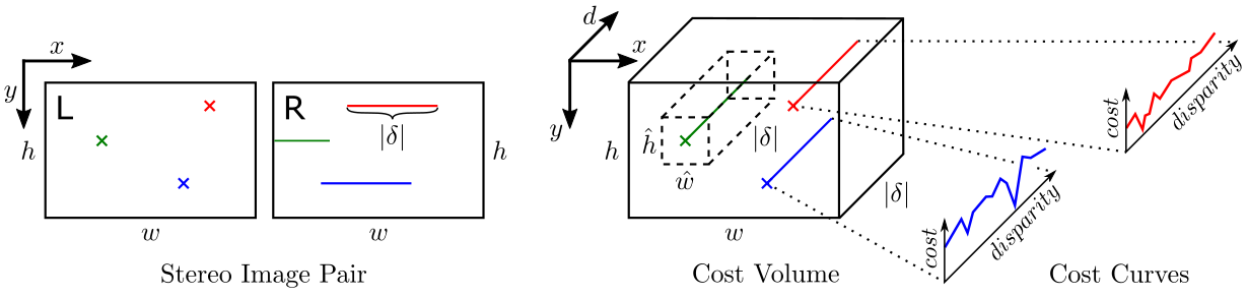


Figure 2.2: Stereo image pair, cost volume and cost curves. Given an epipolar rectified stereo image pair, each of size $\hat{w} \times \hat{h}$, a cost volume of size $\hat{w} \times \hat{h} \times |\delta|$ refers to the left image in this work, where the size in d direction is defined by the considered disparity range $|\delta|$. The cost curve of a pixel in the left image encodes the matching costs of potential correspondences in the right image along the epipolar line and within the specified disparity range $|\delta|$. A cost volume extract of size $\hat{w} \times \hat{h} \times |\delta|$ consists of multiple such cost curves corresponding to pixels located within a local neighbourhood around a pixel of interest in the left image (Mehltretter and Heipke, 2021).

estimates, for example, a clearly identifiable global minimum in the cost curve usually indicates a reliable disparity estimate. One example showing the uncertainty about the disparity estimation is illustrated in Figure 2.1d.

However, the cost curves typically do not provide enough information to detect local outliers. In this context, the cost volume extracts of size $\hat{w} \times \hat{h} \times |\delta|$, consisting of multiple cost curves corresponding to pixels located within a local neighbourhood $\hat{w} \times \hat{h}$ around a pixel of interest in the left image, are processed in this work using the architecture described in Section 2.3.

2.2 Cost Computation

To produce cost volumes that are subsequently processed by the architecture described in the following Section 2.3 and to investigate the generalisation property of our proposed uncertainty models, two different common local stereo algorithms focusing on the matching cost computation,

127	127	129
126	128	129
127	131	A

(a) Intensities of a 3×3 region of an image.

1	1	0
1		0
1	0	a

(b) Set of intensity comparisons.

Table 2.1: Illustration of the census transform.

including conventional Census-based block matching (Zabih and Woodfill, 1994) and deep learning-based MC-CNN fast (Zbontar et al., 2016), are utilised in this work and described in Section 2.2.1 and Section 2.2.2, respectively.

2.2.1 Census Transform

According to Zabih and Woodfill (1994), the *census transform* $R(P)$ maps the local neighborhood $N(P)$ surrounding a pixel P to a bit string representing the set of neighboring pixels whose intensity is less than that of P. Let $\otimes$ denote concatenation, the census transform can then be formed as

$$R(P) = \bigotimes_{Q \in N(P)} \xi(P, Q), \tag{2.2}$$

where the function $\xi(P, Q)$ is set to 1 if the intensity (usually an 8-bit integer) of pixel Q is smaller than that of pixel P. Two pixels of census transformed images are compared for similarity using the Hamming distance, i.e. the number of bits that differ in the two bit strings.

Since the census transform relies solely upon the set of comparisons, this method is more stable in contrast to parametric measures, which is illustrated in Table 2.1. While the mean and variance of a 3×3 region (c.f. Tab. 2.1a) vary from 114 to 142 and from 2 to 1823, respectively, as A ranges from 0 to 255, the census transform simply results in the bits $\{1,1,0,1,0,1,0,a\}$ with a being 1 if $A < 128$ and 0 otherwise.

2.2.2 MC-CNN fast

MC-CNN fast is a deep learning-based local method for extracting cost volume from a rectified image pair, which is proposed by Zbontar et al. (2016). Taking a pair of small image patches as input, the MC-CNN fast architecture, consisting of two shared-weight sub-networks, extracts two vectors capturing the properties of the input patches and outputs a cosine similarity measure of these two vectors. Figure 2.3 provides an overview of this architecture.

The matching cost for the centre pixel of the left input patch is then initialized by taking the negative of network output, i.e. the similarity score. As the MC-CNN fast architecture computes

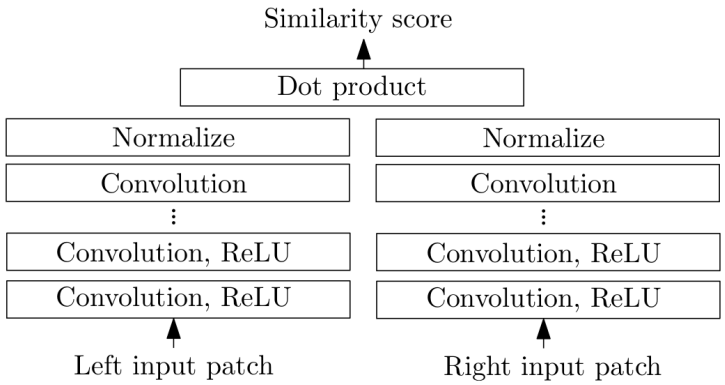


Figure 2.3: Overview of the MC-CNN fast architecture (Zbontar et al., 2016). The two sub-networks consist of a number of convolutional layers with rectified linear units (ReLU) following all but the last layer. Two feature vectors extracted from each of the two input patches by the sub-networks are then compared using the cosine similarity measure, which is split int two steps: normalization and dot product, to produce the final similarity score.

the cosine similarity score, the matching cost ranges from -1 to 1. To keep the running time manageable, the output of the two sub-networks can be computed for all pixels in a single forward pass by propagating full-resolution images, instead of small image patches. The only layer that is run d times is the dot product of the feature vectors to perform the forward pass for each disparity, where d denotes the maximum disparity of the training samples.

2.3 CVA-Net

To evaluate our proposed probabilistic uncertainty models, the architecture of the *Cost Volume Analysis Network* (CVA-Net) is utilised. This network consists of three major processing steps: neighbourhood fusion, depth processing and uncertainty estimation. An overview of this architecture is shown in Figure 2.4.

As input, the network takes cost volume extracts of size $\hat{w} \times \hat{h} \times |\delta|$. According to Mehltretter and Heipke (2021), the size of the perceptive field is set to $\hat{w} = \hat{h} = 13$ pixels, providing a good trade-off between the amount of information available to the network and the degree of smoothing within the resulting uncertainty map. The depth $|\delta|$ of an extract is set to 192 pixels according to the maximum disparity of the training samples.

Driven by the fact, that including neighbourhood information usually increases the robustness, the first part of the network, the neighbourhood fusion, merges a three-dimensional cost volume extract to a single 1D feature vector using 3D convolutional layers with a fixed kernel size of 3×3×3. In order to derive high-level features characterising the cost volume extract, the resulting 1D feature vector is further processed in the subsequent depth processing part using 3D convolutional layers

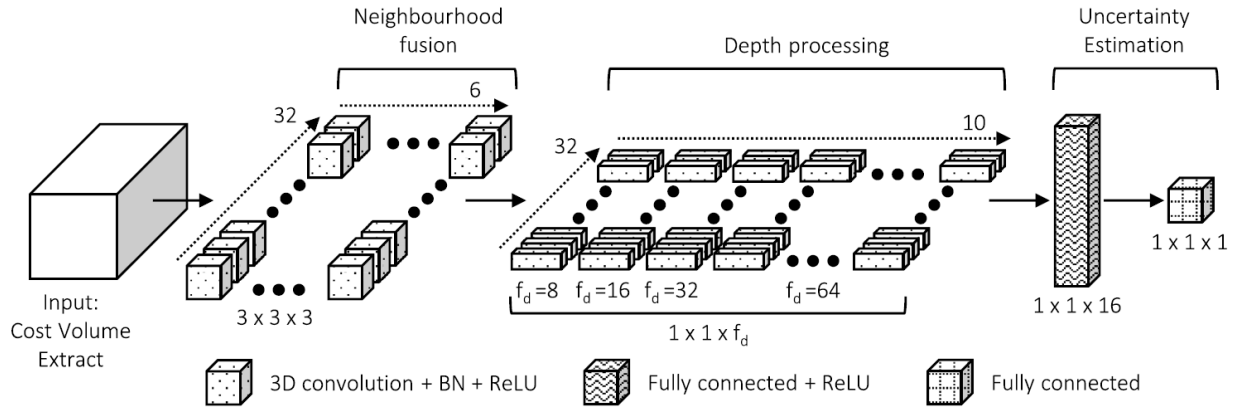


Figure 2.4: Overview of CVA-Net. Consisting of three main processing steps, the network first fuses a cost volume extract into a single 1D feature vector using 3D convolutional layers with a fixed size of 3×3×3. In the second step, the resulting 1D feature vector is further processed using 3D convolutional layers along the disparity axis. The depth of the filter kernels f_d is varying in this part. Based on the extracted features, the fully-connected layers at the end of the network estimate the aleatoric uncertainty pixel by pixel (Mehltretter and Heipke, 2021).

with varying kernel depth f_d. Starting with $f_d = 8$ the value is doubled with every new layer until $f_d = 64$ is reached. This design is optimised empirically by the original authors. Furthermore, zero padding is utilised for all convolutions in this part, to keep the size of the output tensor constant. Based on the extracted features, the fully-connected layers in the last part of the network predict an uncertainty value pixel by pixel. Moreover, to be able to train on image patches while computing an uncertainty map of the full resolution image in a single forward pass at test time, the fully-connected layers are implemented in a convolutional manner. A detailed layer-by-layer definition can be found in Table 2.2.

2.4 Probability Distributions

In this section, four different types of distributions are introduced, i.e., Gaussian distribution in Section 2.4.1, Laplacian distribution in Section 2.4.2, Uniform distribution in Section 2.4.3 and Gaussian mixture distribution in Section 2.4.4. Figure 2.5 shows the first three distribution types with mean μ being 0 and standard deviation σ being 1.

2.4.1 Gaussian Distribution

The Gaussian distribution $x \sim \mathcal{N}(\mu, \sigma^2)$, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the density

Layer	Description	Output Tensor Dimensions
Input	Cost Volume Extract	13×13×192
Neighbourhood Fusion		
1	3D conv., 3×3×3, 32 filters	11×11×190
2	3D conv., 3×3×3, 32 filters	9×9×188
3	3D conv., 3×3×3, 32 filters	7×7×186
4	3D conv., 3×3×3, 32 filters	5×5×184
5	3D conv., 3×3×3, 32 filters	3×3×182
6	3D conv., 3×3×3, 32 filters	1×1×180
Depth Processing		
7	3D conv., 1×1×8, 32 filters, zero pad.	1×1×180
8	3D conv., 1×1×16, 32 filters, zero pad.	1×1×180
9	3D conv., 1×1×32, 32 filters, zero pad.	1×1×180
10-16	3D conv., 1×1×64, 32 filters, zero pad.	1×1×180
Uncertainty Estimation		
17	Fully-connected, 16 nodes, no BN	1×1×16
18	Fully-connected, 1 node, linear act., no BN	1×1×1

Table 2.2: Summary of the CVA-Net architecture. Unless otherwise specified, each layer is followed by batch normalisation (BN) and a ReLU non-linearity (Mehltretter and Heipke, 2021).

function of the Gaussian distribution can be written in the form

$$p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big), \tag{2.3}$$

where μ is the mean and σ is the standard deviation. The simplest case of a Gaussian distribution is known as the *standard Gaussian distribution* with $\mu = 0$ and $\sigma = 1$ (c.f. Fig. 2.5).

The standard deviation σ describes the width of the Gaussian distribution. It holds approximately:

- About 68.27% of values drawn from a Gaussian distribution are within the interval $\mu \pm \sigma$.
- About 95.45% of values drawn from a Gaussian distribution are within the interval $\mu \pm 2\sigma$.
- About 99.73% of values drawn from a Gaussian distribution are within the interval $\mu \pm 3\sigma$.

2.4.2 Laplacian Distribution

The Laplacian distribution $x \sim \mathcal{L}(\mu, b)$ is a continuous distribution with μ being the mean and b being the scale parameter. In contrast to Gaussian distribution, the Laplacian density is expressed

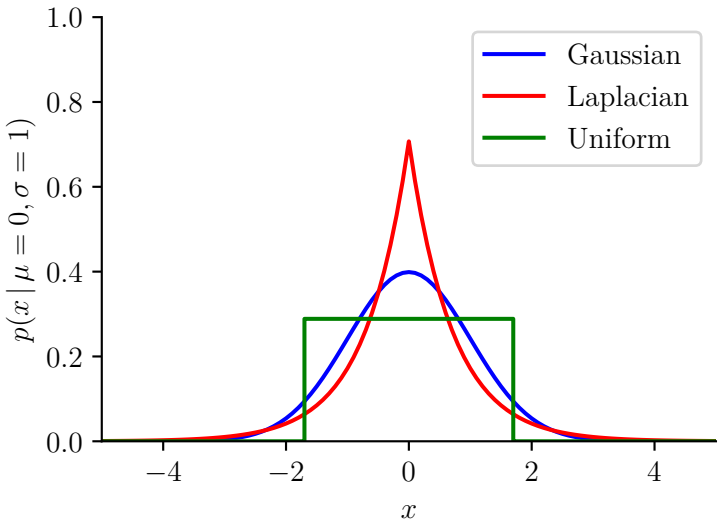


Figure 2.5: Probability Distributions for a single variable x with mean $\mu = 0$ and standard deviation $\sigma = 1$.

in terms of the absolute difference from the mean μ. The probability density function is defined as

$$p(x \mid \mu, b) = \frac{1}{2b} \exp\Big(- \frac{|x - \mu|}{b} \Big). \tag{2.4}$$

Considering the relationship $\sigma = \sqrt{2}\, b$ between the standard deviation σ and the variable b of a Laplacian distribution, we substitute b with σ in the probability density function

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2}\sigma} \exp\Big(- \frac{\sqrt{2}\,|x - \mu|}{\sigma} \Big). \tag{2.5}$$

In this way, the probability density function, i.e. the likelihood can be described directly by the mean μ and the standard deviation σ. A Laplacian distribution with $\mu = 0$ and $\sigma = 1$ is illustrated in Figure 2.5.

Consequently, the negative log likelihood is then formulated as

$$-\log p(x \mid \mu, \sigma) = \frac{\sqrt{2}\,|x - \mu|}{\sigma} + \log \sigma + \underbrace{\log \sqrt{2}}_{\text{constant}}. \tag{2.6}$$

2.4.3 Uniform Distribution

The Uniform distribution $x \sim \mathcal{U}(a, b)$ has a constant probability density on an interval $[a, b]$. The probability density function is defined as

$$p(x \mid a, b) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

The mean μ and the standard deviation σ can then be derived from the probability density function using the following equations:

$$\mu = E[x] = \int_{-\infty}^{\infty} f(x) \cdot x \; dx = \int_{a}^{b} \frac{1}{b-a} \cdot x \; dx = \frac{a+b}{2}, \tag{2.8}$$

$$\begin{aligned} \sigma &= \sqrt{E[(x - E[x])^2]} = \sqrt{E[x^2] - (E[x])^2} \\ &= \sqrt{\int_{a}^{b} \frac{1}{b-a} \cdot x^2 \; dx - \left(\frac{a+b}{2}\right)^2} = \frac{(b-a)}{2\sqrt{3}}. \end{aligned} \tag{2.9}$$

Similar to Laplacian distribution described in Section 2.4.2, the probability density function of the Uniform distribution can also be represented by the mean μ and the standard deviation σ:

$$p(x \mid \mu, \sigma) = \begin{cases} \frac{1}{2\sqrt{3}\,\sigma} & \text{if } -\sqrt{3}\,\sigma \leq x - \mu \leq \sqrt{3}\,\sigma \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

A Uniform distribution with $\mu = 0$ and $\sigma = 1$ is shown in Figure 2.5.

2.4.4 Gaussian Mixture Distribution

While the Gaussian and Laplacian distribution are widely used and have some important analytical properties, they suffer from significant limitations when they come to modelling real datasets, where multimodality occurs. In this case, a linear superposition of multiple unimodal distributions, such as Gaussian distributions, gives a better characterization of the dataset.

Such a Gaussian mixture distribution takes linear combinations of multiple (K) Gaussian distributions (see Sec. 2.4.1) and can give rise to very complex densities (c.f. Fig. 2.6). By using a sufficient number of Gaussian distributions, and by adjusting their means μ_k and variances σ_k^2 as well as the coefficients ϕ_k in the linear combination, almost any continuous density can be approximated to arbitrary accuracy (Bishop, 2006). The form of a superposition of K Gaussian densities

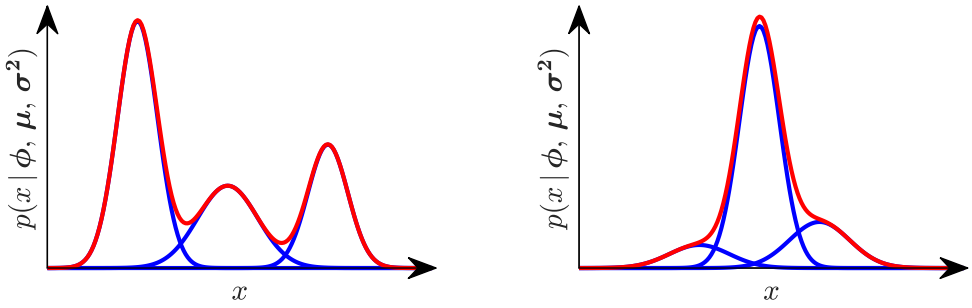


Figure 2.6: Example of Gaussian mixture distributions showing each with three Gaussian components (each scaled by a mixture coefficient) in blue and their sum in red. While three modes are shown on the left, the Gaussian mixture distribution on the right has only one single mode.

is considered as

$$p(x \mid \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = \sum_{k=1}^{K} \phi_k \cdot \mathcal{N}(x \mid \mu_k, \sigma_k^2), \tag{2.11}$$

where $\boldsymbol{\phi}$, $\boldsymbol{\mu}$, and $\boldsymbol{\sigma^2}$ denote the K-dimensional vector of mixture coefficients, means and variances, respectively. The negative log likelihood is then formulated as

$$-\log p(x \mid \boldsymbol{\phi}, \boldsymbol{\mu}, \boldsymbol{\sigma^2}) = -\log\left(\sum_{k=1}^{K} \phi_k \cdot \mathcal{N}(x \mid \mu_k, \sigma_k^2) \right). \tag{2.12}$$

According to Carreira-Perpinan (2000), the mean μ and variance σ^2 of the Gaussian mixture distribution can be computed using the following formulations:

$$\mu = \mathrm{E}[x] = \sum_{k=1}^{K} \phi_k \, \mu_k \tag{2.13}$$

$$\begin{aligned} \sigma^2 &= \mathrm{E}[(x - \mathrm{E}[x])^2] = \mathrm{E}[x^2] - (\mathrm{E}[x])^2 \\ &= \left(\sum_{k=1}^{K} \phi_k \, \mathrm{E}_{p(x|k)}[x^2] \right) - \mu^2 \\ &= \left(\sum_{k=1}^{K} \phi_k \, (\sigma_k^2 + (\mathrm{E}_{p(x|k)}[x])^2) \right) - \mu^2 \\ &= \left(\sum_{k=1}^{K} \phi_k \, (\sigma_k^2 + \mu_k^2) \right) - \mu^2, \end{aligned} \tag{2.14}$$

where $p(x \mid k)$ denotes the conditional probability of variable x for the k-th Gaussian component. Note that these results are valid for any mixture distributions, not necessarily of Gaussians.

3 Related Works

In this chapter, the related works relevant for this thesis are described. First, the two popular aleatoric uncertainty estimation methods in the context of dense stereo matching, i.e. confidence-based and probabilistic-based methods, are briefly outlined in Section 3.1. Based on the uncertainties estimated by these methods, a variety of metrics are introduced in Section 3.2 to evaluate different uncertainty models.

3.1 Uncertainty Estimation

The uncertainty information has many applications in the context of dense stereo matching: Such information is used to reliably reject outliers that arise in multi-view stereo (Vogiatzis and Hernandez, 2011), or to de-noise the estimated depth maps for monocular dense reconstruction (Pizzoli et al., 2014). Furthermore, multiple stereo matching algorithms (Poggi and Mattoccia, 2016a), different cost functions (Mehltretter et al., 2018) or cost volumes (Sun et al., 2017) are weighted aggregated using the information regarding uncertainty.

Together with these applications, many approaches have been proposed in recent years to estimate the uncertainty of disparity assignments. At first, hand-crafted metrics were designed to quantify aleatoric uncertainty, such as Peak Ratio (PKR), which is designed based on the properties of the cost curve, and Left Right Difference (LRD), which considers the consistency between the left and the right disparity maps. A good overview of the commonly used hand-crafted metrics is given in (Hu and Mordohai, 2012).

Similar to other computer vision fields, more and more approaches based on deep learning (Poggi and Mattoccia, 2016b; Mehltretter and Heipke, 2019; Kendall and Gal, 2017) and other machine learning techniques (Sun et al., 2017; Batsos et al., 2018) have been proposed in the literature. While a majority of these deep learning-based uncertainty estimation methods, operate on extracted patches from disparity maps only (Poggi and Mattoccia, 2016b) or additionally take the RGB reference image into account (Fu et al., 2019), Mehltretter and Heipke (2019) utilise the information contained in the 3D cost volume. Such a cost volume is an intermediate representation present in most dense stereo matching algorithms which typically contains additional information compared to disparity maps. Benefiting from this additional information, cost volume-based approaches have

demonstrated to allow a more accurate estimation of the uncertainty.

While plenty of literature exists focusing on various types of features, modelling the aleatoric uncertainty has received significantly less attention. Among these uncertainty estimation methods, the confidence-based strategy is most popular (Fu et al., 2019; Tosi et al., 2018). Driven by the fact that, in contrast to depth, ground truth is typically not available for the associated uncertainty, uncertainty prediction has to be learned implicitly. For this purpose, confidence-based methods predict a score per pixel between zero and one, representing the trust on the corresponding depth estimate, and thus, can be learned as a binary class probability of a depth estimate being correct or incorrect.

Another strategy for aleatoric uncertainty estimation that recently has received increasing attention is the probabilistic one. In contrast to confidence-based methods, probabilistic methods assume a probabilistic model, commonly in form of a certain probability distribution, for the aleatoric uncertainty (Kendall and Gal, 2017; Mehltretter and Heipke, 2021). In this context, the depth and the associated uncertainty are considered as the mean and variance (or standard deviation) of the presumed model, respectively. With the reference depth being used as observation, these models can be trained with the objective of maximising the likelihood. Since this approach is based on the real error distribution, the uncertainty can be additionally quantified in pixels or in metric units.

In all of these probabilistic methods, only a unimodal distribution is used to model the aleatoric uncertainty, considering the error distribution as a Gaussian (Kendall and Gal, 2017) or a Laplacian distribution (Poggi et al., 2020; Mehltretter and Heipke, 2021). However, this is not always the case in the context of dense stereo matching, especially in real-world scenarios this assumption is violated by outlier measurements or by commonly challenging regions, such as occlusion, texture-less regions and depth discontinuities. Simply neglecting these cases leads to an over-simplification that may reduce the accuracy of the estimated aleatoric uncertainty.

3.2 Assessment of Predictive Uncertainty

While plenty of widely used metrics are utilised for the evaluation of disparity estimates, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) etc., there are significantly fewer standard metrics to evaluate the predictive uncertainty. Qualitative evaluations are commonly performed visually in comparison to images of the estimated uncertainty. In terms of quantitative evaluation in the literature of dense stereo matching, the most common metric is Area Under the Curve (AUC) (Hu and Mordohai, 2012) which measures how much the estimated uncertainty correlates with the true error rate. If pixel correspondences with the lowest uncertainty values are selected gradually, the smallest possible AUC for a given disparity map can then be obtained.

One limitation of AUC metric is that it only evaluates the rankings of the estimated uncertainty values, while it simply neglects the actual magnitude of the estimated uncertainty and thus also the relation between this estimates and the real disparity error. To overcome this limitation, Mehltretter and Heipke (2021) use the Pearson correlation coefficient between the absolute disparity error and the estimated uncertainty to evaluate the quantified uncertainty. However, the correlation coefficient is known to be very sensitive to the presence of outliers (Abdullah, 1990), which can occur in both the uncertainty estimates and the datasets.

In the forecasting literature, Gneiting et al. (2007) propose to evaluate a predictive uncertainty forecast in terms of calibration and sharpness. Calibration refers to the statistical consistency between the predictive distributions and the observations. Sharpness refers to the concentration of the predictive distributions. In terms of forecasting, calibration means that whenever a forecaster assigns a probability of 0.8 to an event, that event should occur about 80% of the time. Calibration alone is not sufficient to ensure a useful forecast, the predictions also need to be sharp, i.e. close to zero or one. More recently, Kuleshov et al. (2018) propose a method to recalibrate regression algorithms so that the uncertainty estimates lie in the expected confidence interval. They also propose metrics to assess both calibration and sharpness of a predictive distribution. Regarding regression, calibration means that the empirical and the predicted cumulative density functions (CDFs) should match. Meanwhile, the confidence intervals should all be as tight as possible around a single value. As both calibration and sharpness require prior knowledge of the CDF for the assumed distribution (Kuleshov et al., 2018), they are unfortunately not applicable for our proposed GMM and mixture uncertainty models (see Sec. 4.3.3 and 4.3.4) due to the varying mixture coefficients.

While Kuleshov et al. (2018) consider calibration and sharpness separately, Moukari et al. (2019) propose a metric Normalized Mean Rescaled Confidence Interval (n-MeRCI) to evaluate the correlation between predictive uncertainty and true error, which implicitly considers them together independent of the assumed error distribution. Moreover, the n-MeRCI is only performed on a certain number of the data, which avoids a few data points to strongly impact the whole measure and therefore increases the robustness. However, this requires the prior knowledge of the inlier rate of predictions. In addition, the way of this metric to define the inliers is sensitive to data points with small errors, shown in Section 6.2.

4 Methodology

To adjust the probabilistic strategy of aleatoric uncertainty estimation in the context of dense stereo matching to better fit to the characteristics of real-world scenarios, we propose three novel uncertainty models based on three different assumptions, i.e. geometry-aware, GMM and mixture uncertainty models, which are discussed in detail in Section 4.3. For the purpose of a fair comparison of different uncertainty models, i.e. avoiding the impacts coming from the disparity estimation process or different network architectures used to carried out the task of uncertainty prediction, we test all variants using an adaptation of a common neural network architecture (briefly outlined in Section 4.2). Information regarding the input of the network architecture and problem statement is shown in Section 4.1.

4.1 Input and Problem Statement

To predict pixel-wise aleatoric uncertainty during evaluation procedure, we feed existing cost volumes to a trained CNN. These cost volumes are the result of the cost computation step of an arbitrary dense stereo matching approach carried out on epipolar rectified image pairs. Contrary to evaluation procedure, the disparity maps previously computed using the "winner takes all" strategy based on these cost volumes and the ground truth disparities are additionally required for the defined loss functions during training. Furthermore, the binary region masks (see Sec. 5.2) are necessary to allow a separate modeling of aleatoric uncertainty in different regions for our proposed geometry-aware uncertainty model described in Section 4.3.2. While the uncertainty values are inferred in all uncertainty models described in Section 4.3, the models based on mixture distributions, i.e. the GMM and the mixture models, predict further the information regarding each mixture component.

4.2 Basic Architecture

Showing convincing results for the task of aleatoric uncertainty estimation in (Mehltretter and Heipke, 2019, 2021), the architecture of the *Cost Volume Analysis Network* (CVA-Net, c.f. Sec. 2.3) is utilised to evaluate our subsequently proposed probabilistic uncertainty models. Same as in the

Layer	Description	Output Tensor Dimensions
Input	Cost Volume Extract	13×13×192
Neighbourhood Fusion		
1	3D conv., 5×5×5, 32 filters	9×9×188
2	3D conv., 5×5×5, 32 filters	5×5×184
3	3D conv., 5×5×5, 32 filters	1×1×180
Depth Processing		
4	3D conv., 1×1×8, 32 filters, zero pad.	1×1×180
5	3D conv., 1×1×16, 32 filters, zero pad.	1×1×180
6	3D conv., 1×1×32, 32 filters, zero pad.	1×1×180
7-13	3D conv., 1×1×64, 32 filters, zero pad.	1×1×180
Uncertainty Estimation		
14	Global average pooling, linear act., no BN	1×1×32
15	3D conv., 1×1×1, 1 filter, linear act., no BN	1×1×1

Table 4.1: Summary of the modified CVA-Net architecture. The modifications and their consequences are marked in blue and green, respectively. Unless otherwise specified, each layer is followed by batch normalisation (BN) and a ReLU non-linearity. (Adapted from (Mehltretter and Heipke, 2021).)

original architecture, the size of the input cost volume is set to 13×13×192, where the size of the perceptive field, i.e. 13×13, is chosen considering the trade-off between the amount of information available to the network and the degree of smoothing within the resulting uncertainty map. The depth of 192 pixels is set according to the maximum disparity of the training samples.

As a result of preliminary experiments, the kernel size of the 3D convolutional filters in the first part, the neighbourhood fusion, is changed from 3×3×3 to 5×5×5, to reduce the training time while maintaining comparable results. Consequently, every convolutional layer within this part reduces the extends of the feature tensor by 4 in every direction, reducing the number of layers in this part from 5 to 3. Furthermore, the amount of necessary floating point operations (FLOPs) in the forward pass of the network is reduced by about 20%. Due to the decrease of the FLOPs and the number of sequential layers in the first part, the training procedure of the modified network is accelerated by about 60%.

After observing comparable or in some cases even superior results, the first fully-connected layer is substituted by a global average pooling (GAP) layer (Lin et al., 2014). The depth of the output tensor of this layer is therefore changed to 32, corresponding to the number of filters in the previous layer in the depth processing part. In such a way, the number of trainable parameters from the original fully-connected layer can be reduced and consequently, the potential of over-fitting to the

training data is reduced. In addition, the GAP layer allows the network to flexibly process cost volumes with different depths without the need for re-training, since the GAP layer is independent of the output dimensions of the previous layer. A detailed layer-by-layer definition of the modified architecture can be found in Table 4.1.

4.3 Probabilistic Uncertainty Models

Disparity estimation from stereo images is commonly learned in a supervised manner. However, reference data for the associated uncertainty is typically not available. To nevertheless be able to learn the task of uncertainty estimation, it is common to assume a specific uncertainty model, allowing to implicitly learn uncertainty from the deviations between estimated and ground truth disparity. First, the Laplacian model, considering the real error to be unimodally distributed, is reviewed in Section 4.3.1, before three novel probabilistic models are proposed in Sections 4.3.2, 4.3.3 and 4.3.4, taking challenging regions of real-world scenarios and outlier measurements explicitly into account. Besides different loss functions, also minor adjustments of the final network layer may be necessary to allow the prediction of varying numbers and types of values used to parameterise these uncertainty models, explained in detail in the respective paragraphs.

4.3.1 Laplacian Model

Under the assumption that the uncertainty contained in the data can be described with a specific probability distribution, the parameters of this distribution can be inferred by maximising the likelihood (Kendall and Gal, 2017). Similar to (Mehltretter and Heipke, 2021), we consider the disparity d to be estimated in advance and the predicted aleatoric uncertainty σ as the mean and the standard deviation used to parameterise a Laplacian distribution. In this context, the ground truth disparity $\hat{d}$ is used as observation. In this way, aleatoric uncertainty can be learned without the need for a reference of the uncertainty. By formulating the objective of this method as the negative log likelihood of the Laplacian distribution as,

$$-\log p(\hat{d}_i \,|\, d_i) \propto \frac{\sqrt{2}}{\sigma_i}|d_i - \hat{d}_i| + \log(\sigma_i), \tag{4.1}$$

where i is the index of a sample in a mini-batch processed in one forward pass, we enable the use of optimisation techniques commonly employed to train neural networks. To make the training procedure numerically more stable and to prevent the loss function from being divided by zero, $s = \log(\sigma)$ is substituted in the loss function to predict the log standard deviation, as proposed by

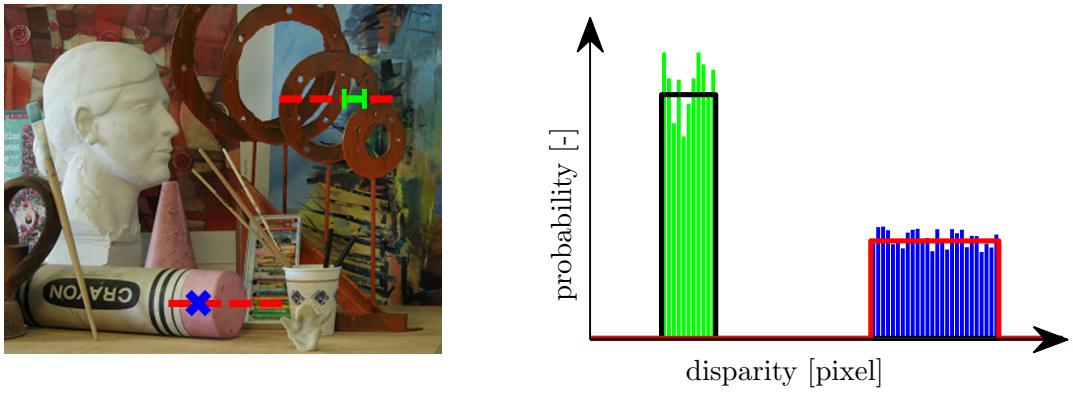


Figure 4.1: Illustration of the underlying assumptions of the proposed geometry-aware model. On the left, the right image of a stereo pair with a correct point match in a texture-less region and matching area in an occluded region are shown in blue and green, respectively. The considered search space is highlighted in red. On the right, the corresponding probability distributions over the considered disparity range are shown. The histograms of the two cases, shown in blue and green, are approximated by Uniform distributions, shown in red and black.

Kendall and Gal (2017). Finally, the loss function of the Laplacian model is defined as:

$$L_{\text{Laplacian}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sqrt{2}}{\exp(s_i)}|d_i - \hat{d}_i| + s_i, \tag{4.2}$$

where N is the number of training samples with known ground truth disparity.

This model has already proven its suitability in the context of aleatoric uncertainty estimation for dense stereo matching (Mehltretter and Heipke, 2021) and will therefore be used as baseline in this work.

4.3.2 Geometry-aware Model

The previously described variant models the uncertainty contained in the data using a Laplacian distribution. Assuming a unimodal error distribution for each pixel, the mode of such a distribution is designed to match with a unique and distinct global minimum in the corresponding cost curve over the whole disparity range. However, this assumption is often not valid for all pixels in an image.

In the context of dense stereo matching, a commonly challenging scenario is the presence of texture-less regions. A typical behaviour for pixel in such non- or weakly textured regions is the occurrence of a wide and flat minimum in the cost curve, causing a wide and flat maximum in the corresponding probability density function. This maximum can be described by a range of (almost) uniform probability with a length corresponding to the width of the non-textured region (c.f. Fig. 4.1).

Similar to texture-less regions, also pixels in occluded regions, where some parts of a scene are not visible in one of the images of a stereo pair, may cause deviations from the assumption of a unimodal probability density function. Because pixels in these regions do not have a correspondence, occluded pixels are often characterised by the absence of a distinct global minimum, but multiple local minima within an individual disparity interval in the corresponding cost curve. For this model, we assume that all disparities in such an interval have the same probability to be correct. Note that this is a strong simplification, since a uniform probability density function can not represent such local minima of a cost curve properly (c.f. Fig. 4.1). To overcome this limitation, a multimodal probability density function, for example, described as Gaussian mixture model, could be employed to describe the characteristics contained in these regions.

As both, texture-less and occluded regions, are assumed to be uniformly distributed in specific intervals in the probability density function, a Laplacian distribution is not sufficient to model the uncertainty in these areas. Consequently, we use a Uniform distribution with a predictive interval in our geometry-aware variant to model the uncertainty in these "hard" regions. Equal to the Laplacian variant described in Section 4.3.1, we consider the estimated disparity d as the mean and as the centre of an interval with Uniform distribution, respectively. To achieve maximal probability, the ground truth disparity $\hat{d}$ needs to be included in the interval $[d-r, d+r]$, where r denotes the half interval length, without motivating the network to predict unreasonable large intervals. Thus, we maximise the likelihood with respect to the ground truth disparity $\hat{d}$ in these "hard" regions, by minimising the difference between the predictive half interval length r and the absolute error of disparity estimation $|d - \hat{d}|$.

Considering the relationship $r = \sqrt{3}\,\sigma$ between the half interval length r and the standard deviation σ of a Uniform distribution, we use the Huber loss (Huber, 1981), which combines the advantages of L1-loss (steady gradients for large values x) and L2-loss (less oscillation during updates when x is small), to state the optimisation objective for the "hard" regions:

$$L_{\mathcal{U}} = \begin{cases} 0.5x^2 & \text{if } |x| \leq \gamma \\ \gamma|x| - 0.5\gamma^2 & \text{otherwise,} \end{cases} \tag{4.3}$$

in which $x = |d_i - \hat{d}_i| - \sqrt{3}\,\sigma_i$. In this work, we set γ to 1.

The cost curves of pixels located in none of the previously mentioned regions are usually characterised by a distinct global minimum. This allows to reliably determine the correct pixel correspondence in most cases and thus leads to a lower uncertainty. Consequently, equal to the first variant, we use the Laplacian distribution to describe the uncertainty for pixels located in these "good" regions:

$$L_{\mathcal{L}} = \frac{\sqrt{2}}{\exp(s_i)}|d_i - \hat{d}_i| + s_i. \tag{4.4}$$

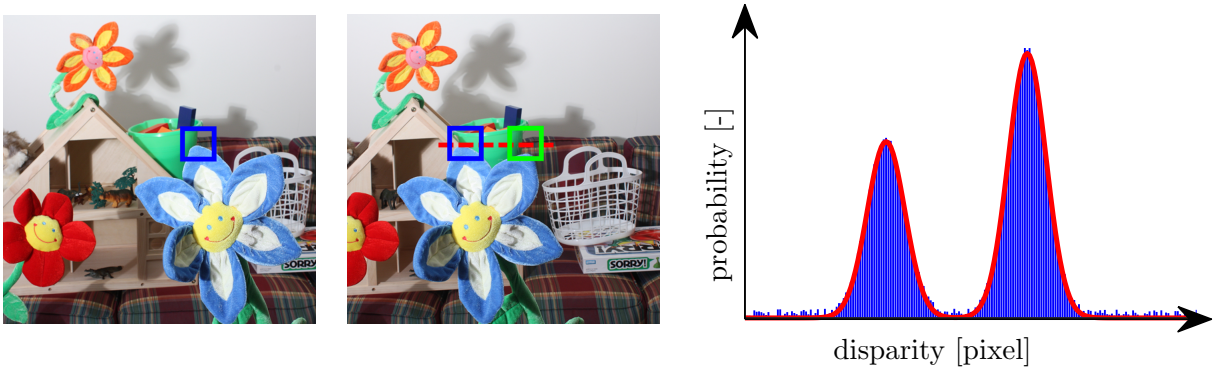


Figure 4.2: A scenario at depth discontinuities. From left to middle, the left and right image of a stereo pair are shown. In the left image, the neighbourhood window of a pixel at depth discontinuities is marked with a blue rectangle. In the right image, besides the true correspondence marked with a blue rectangle, an additional potential correspondence marked with a green rectangle can be seen along the considered search space (highlighted with a red dashed line). This results in a bimodal probability distribution in the considered disparity interval, which is shown on the right. The histograms of disparities, shown in blue, are approximated by a Gaussian mixture model with two components, shown in red. Note that this is a striking example, which may differs from the actual probability histograms.

For the purpose of consistency, we also substitute $s = \log(\sigma)$ for the objective of "hard" regions, so that our network is trained to predict the log standard deviation for both types of regions and probability distributions. We therefore define the loss function of our geometry-aware variant as:

$$L_{\text{Geometry}} = \frac{1}{N} \sum_{i=1}^{N} c_i \cdot L_{\mathcal{L}} + (1 - c_i) \cdot L_{\mathcal{U}}, \tag{4.5}$$

where c denotes a binary parameter that is computed according to (Scharstein and Szeliski, 2002) and discussed in detail in Section 5.2. A pixel i gets assigned $c_i = 1$ if it is located in a "good" region and 0 otherwise.

4.3.3 GMM Model

As already mentioned in Section 4.3.2, the aleatoric uncertainty for occluded pixels could not be described properly using a Uniform distribution, since they are often characterised by multiple local minima in the corresponding cost curve. Such local minima lead to multimodality of the corresponding probability density function, which could be represented by a Gaussian mixture model (GMM) (see Sec. 2.4.4).

With the help of GMM, we could also be able to better approximate the uncertainty of pixels at depth discontinuities, also one commonly challenging scenario in the context of dense stereo matching. One example is illustrated in Figure 4.2. The neighbourhood of a pixel at the border of the

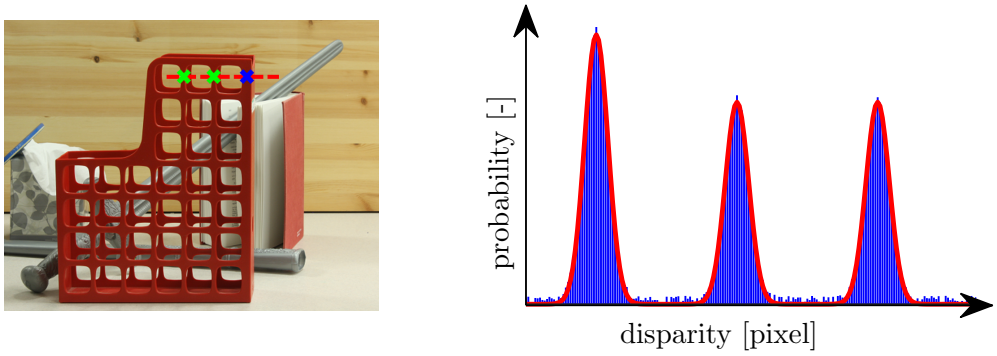


Figure 4.3: A scenario of area with repetitive patterns. On the left, the right image of a stereo pair is shown with a correct point match marked in blue and two additional potential point correspondences marked in green. This results in three modes in the corresponding probability distribution, which are approximated by Gaussian mixture model with three components in red on the right. Note that this is a striking example, which may differs from the actual probability histograms.

flower toy is marked with a blue rectangle in the left image of a stereo pair. Along the epipolar line under consideration of a finite disparity interval, there is no distinct correspondence pixel in the right image, whose neighbourhood is the same as that in the left image, but rather two potential correspondences. In the corresponding probability distribution, this scenario results in a bimodal distribution and thus, can be modeled by a Gaussian mixture model with two components. In general, the number of modes in the probability curve depends on the scene in the local neighbourhood of the pixel of interest, e.g. if the neighbourhood contains not only foreground and background objects, but several objects with different depths, even more than two modes are possible.

In the context of dense stereo matching, the multimodality of the probability distribution may also occur in areas with repetitive patterns (c.f. Fig. 4.3). In these areas, a pixel in the left image of a stereo pair usually has multiple potential correspondences due to the presence of repetitive patterns. Contrary to a unimodal distribution (see Sec. 4.3.1), the multimodality can be better approximated by a Gaussian mixture model. One advantage of a Gaussian mixture model against a unimodal distribution is that, a Gaussian mixture model is not only able to model multimodality, but also unimodality by assigning the mixture coefficients, means and variances of Gaussian components properly (c.f. Fig. 2.6).

Unlike in Section 4.3.2, where the uncertainty is explicitly and strictly treated differently by the defined region types, in this method, this is learned implicitly from the data by assuming a Gaussian mixture model with K components. To learn the proper parameters of the Gaussian mixture model, i.e., mixture coefficients $\boldsymbol{\phi}$, means $\boldsymbol{\mu}$ and variances $\boldsymbol{\sigma^2}$, we utilise the same optimization strategy as in Section 4.3.1 and 4.3.2, i.e., likelihood maximisation. To make the training procedure numerically more stable and to prevent the loss function from being divided by zero, we substitute $\mathbf{s} = \log(\boldsymbol{\sigma^2})$ in the loss function. Note that this is slightly different from Section 4.3.1 and 4.3.2, the log variance

is predicted in stead of log standard deviation in this method. Thus, the negative log likelihood (NLL) loss term becomes:

$$\begin{aligned} L_{\text{GMM-NLL}} &= -\log\left(\sum_{k=1}^{K} \phi_{ik} \cdot \mathcal{N}(d_i \,|\, \mu_{ik}, \sigma_{ik})\right) \\ &= -\log\left(\sum_{k=1}^{K} \frac{\phi_{ik}}{\sqrt{2\pi \cdot \exp(s_{ik})}} \exp\left(\frac{-(\hat{d}_i - \mu_{ik})^2}{2 \cdot \exp(s_{ik})}\right)\right). \end{aligned} \tag{4.6}$$

As a minus sign occurs in the last exponential term, this part could be very close to zero. Consequently, the underflow problem could arise, i.e. the result of this part could be smaller in magnitude (that is, closer to zero) than the computer can actually represent. In this case, the computer would interpret this number as zero. While this error is imperceptible in common situations, it results in numerical issues when a logarithm function is inserted between operations. To eliminate the numerically unstable behaviour caused by this problem, we utilise the *log-sum-exp* trick:

$$\log \sum_{k=1}^{K} \exp(x_k) = \log \sum_{k=1}^{K} \exp(x_k) \cdot \exp(X - X) = X + \log \sum_{k=1}^{K} \exp(x_k - X), \tag{4.7}$$

with $X = \max(x_k)$ for $k \in \{1, ..., K\}$. In this way, we guarantee that the term $\sum_{k=1}^{K} \exp(x_k - X)$ is bigger than 1 and the underflow problem is therefore solved. To enable the utilisation of the log-sum-exp trick, the NLL loss term muss be converted to the corresponding log-sum-exp form:

$$L_{\text{GMM-NLL}} = -\log\left(\sum_{k=1}^{K} \exp\left(\log \phi_{ik} - \frac{\log 2\pi}{2} - \frac{s_{ik}}{2} - \frac{(\hat{d}_i - \mu_{ik})^2}{2 \cdot \exp(s_{ik})}\right)\right). \tag{4.8}$$

To avoid that the learned GMM fits the observation perfectly, i.e. the ground truth disparity $\hat{d}$, while deviate from the estimated disparity d severely, we introduce a mean shift error as a penalty term to ensure that the mean of GMM is near the estimated disparity:

$$L_m = (d_i - \mu_i)^2, \tag{4.9}$$

where μ_i denotes the mean of K Gaussian components for the i-th sample and can be computed using $\mu_i = \sum_{k=1}^{K} \phi_{ik} \cdot \mu_{ik}$ (see Sec. 2.4.4). Finally, the loss function of our GMM variant is defined as:

$$L_{\text{GMM}} = \frac{1}{N} \sum_{i=1}^{N} (L_{\text{GMM-NLL}} + \beta \cdot L_m), \tag{4.10}$$

where β is a hyper parameter to control the penalty term and is set to 1 in this work based on preliminary experiments. As the mixture coefficient ϕ_{ik}, mean μ_{ik} and log variance s_{ik} are predicted for each Gaussian component k in this method, the last layer in Table 4.1 is substituted by three

layers with K 1×1×1 filters each, where K is set to 3 in this work according to the results of preliminary experiments.

4.3.4 Laplacian Uniform Mixture Model

While commonly challenging regions in the context of dense stereo matching are explicitly addressed in our geometry-aware model, we treat aleatoric uncertainty from the perspective of measurement reliability in our mixture model presented in this section. According to Vogiatzis and Hernandez (2011) and Pizzoli et al. (2014), a depth sensor produces two types of measurement: (1) a good measurement that is unimodally distributed around the correct depth or (2) an outlier measurement that is drawn from a Uniform distribution defined on a certain interval. Similar to the geometry-aware model (see Sec. 4.3.2), we assume a Laplacian distribution for good measurements and a Uniform distribution with a predictive interval for outlier measurements. Thus, from the perspective of measurement reliability, aleatoric uncertainty can be described as a mixture of a Laplacian and a Uniform distribution assigned a probability α and $1 - \alpha$, respectively.

Using the same optimisation objectives for both types of distributions as in Section 4.3.2, the loss function of our mixture model is defined as:

$$L_{\text{Mixture}} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \cdot L_{\mathcal{L}} + (1 - \alpha_i) \cdot L_{\mathcal{U}}, \tag{4.11}$$

where the inlier probability α is predicted by the network, together with the log standard deviations of the Laplacian distribution $s_{\mathcal{L}}$ and the Uniform distribution $s_{\mathcal{U}}$. The number of output nodes in the network architecture is therefore increased to three in this variant. To impose positive values and to ensure that the inlier and outlier probabilities sum up to one, a softmax transformation is applied to the α-node.

In summary, this variant can be viewed as an extension of our geometry-aware model. In case that the inlier probability α is predicted as 1 and 0 for "good" and "hard" regions respectively, both variants are equal (c.f. Eq. 4.5 and Eq. 4.11). However, due to the differences between the binary region parameter c in the geometry-aware model and the inlier probability α in the mixture model, the meaning of both variants is rather different.

5 Experimental Setup

In order to investigate the influence of the different uncertainty models (c.f. Sec. 4.3), we train and evaluate these four models on three different datasets (discussed in detail in Section 5.1) as well as on cost volumes computed by two popular stereo matching methods: Census-based block matching (with a support region size of 5×5) (Zabih and Woodfill, 1994) and MC-CNN fast (Zbontar et al., 2016). For the evaluation, we use three metrics that are described in Section 5.4. To allow a fair comparison, all examined methods have been trained on the same data using the same network architecture, following the procedure described in Section 5.3.

5.1 Datasets

In this work, three datasets, namely KITTI 2012 (Geiger et al., 2012), KITTI 2015 (Menze and Geiger, 2015) and Middlebury v3 (Scharstein et al., 2014) are used for the experiments. The KITTI datasets consist of challenging and varied road scenes captured from a moving vehicle. Ground truth disparity maps for training and evaluation are obtained from LIDAR data with disparities for about 30% of the pixels. In contrast to the KITTI datasets, the Middlebury dataset contains 15 image pairs showing various indoor scenes captured with a static stereo set-up and provides dense and highly accurate ground truth disparity maps based on structured light.

5.2 Binary Region Masks

For the training procedure of the geometry-aware uncertainty model discussed in Section 4.3.2, binary region masks are needed, which indicate with 1 and 0, respectively, whether a pixel is located in a "good" or a "hard" region. Since the ground truth disparity maps of the KITTI datasets are not dense, masks for depth discontinuities cannot be computed accurately. In this work, we therefore consider texture-less regions and occluded areas as "hard" regions with the following definitions:

- Texture-less regions: regions in which the squared horizontal intensity gradient averaged over a 3×3 window is smaller than 4.0 (Scharstein and Szeliski, 2002).

- Occluded areas: occlusion can be determined from ground truth disparity maps directly, but the corresponding masks are also already provided as part of all three datasets.

Pixels located in none of these two types of "hard" regions are in turn labelled as "good".

5.3 Training Procedure

Following the description in (Mehltretter and Heipke, 2021), we train the four uncertainty models on the first 20 training image pairs of the KITTI 2012 dataset (Geiger et al., 2012) and use three additional image pairs for validation. As input for the network, tensors of size $13 \times 13 \times 192$ are extracted from normalised cost volumes corresponding to the left image of each pair, where the depth of 192 pixels is set according to the maximum disparity of the training samples. The values contained in these cost volumes are normalised to $[0, 1]$ using min-max normalisation. Every extract is centred on a pixel with available ground truth disparity. 128 of such extracts are bundled to one mini-batch and fed to the network per forward pass during training.

We initialise the convolutional layers with normal distributions $\mathcal{N}(0, 0.0025)$ and use the Adam optimiser (Kingma and Ba, 2015) with its parameters set to their default values. The learning rate is set to 10^{-4}. To enforce generalisation, dropout (Srivastava et al., 2014) is applied to the global average pooling layer (layer 14, c.f. Tab. 2.2) with a rate of 0.5. The training procedure is stopped, if the validation loss did not decrease in three consecutive epochs and the weights showing the lowest validation loss are used for testing. Thus, we trained the modified CVA-Net for 20, 17, 9 and 19 epochs for the Laplacian, the geometry-aware, the GMM and the mixture model, respectively.

5.4 Metrics

Following Mehltretter and Heipke (2021), we first use the AUC metric (c.f. Sec. 5.4.1) und correlation coefficient (c.f. Sec. 5.4.2) to evaluate the different uncertainty models described in Section 4.3. The AUC metric measures the relative order of the uncertainty values, while the correlation coefficient calculates the linear dependence between the absolute disparity error and the uncertainty estimate. Considering the outliers that could occur in the datasets we are using, we use the normalized mean rescaled confidence interval (n-MeRCI) (Moukari et al., 2019) as an additional metric to evaluate the estimated uncertainty values, introduced in Section 5.4.3.

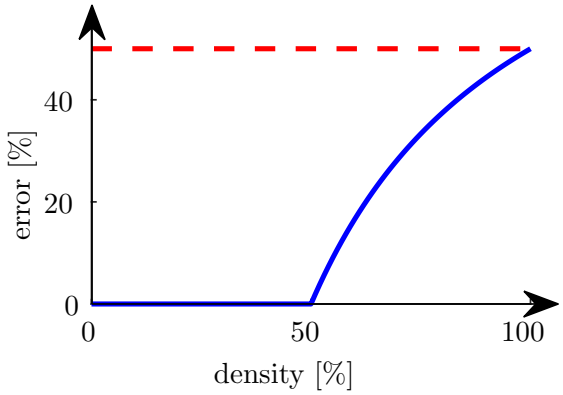


Figure 5.1: An example of AUC calculation for a disparity map with an overall error 50%. The optimal AUC (all correct matches selected first) and AUC for random chance are shown in blue and red, respectively.

5.4.1 AUC

The first metric we use to evaluate the methodology presented in this work, is the Area Under the Curve (AUC), originally proposed by Hu and Mordohai (2012) to evaluate different confidence estimation methods. In this context, a Receiver Operating Characteristic (ROC) curve, for which the AUC is computed, represents the error rate as a function of the percentage of pixels sampled from a disparity map in order of increasing uncertainty. Specifically, for each uncertainty model, we select the top 5% of the pixels with the lowest uncertainty and record the error rate, defined as the percentage of pixels with a disparity error smaller than 3 pixels or 5% of the ground truth disparities (Menze and Geiger, 2015), then repeat for the top 10% and so on. Ideally, all pixels with correct disparity estimations should be selected before all errors, resulting in the smallest possible AUC for a given disparity map. Random selection of pixels produces a flat ROC curve with an AUC equal to the error rate of the disparity map at full density. An illustration of the optimal AUC and the AUC obtained by random chance is shown in Figure 5.1.

Consequently, the optimal AUC depends only on the overall error ϵ of a disparity map:

$$\begin{aligned} AUC_{opt} &= \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p} dp \\ &= \epsilon + (1-\epsilon)\ln(1-\epsilon), \end{aligned} \tag{5.1}$$

where p is the percentage of pixels sampled from a disparity map. The closer the AUC of an uncertainty map gets to the optimal value, the higher the accuracy.

If many pixels share the same uncertainty, consequently, the specified sampling distance may be exceeded significantly, resulting in a high discretization error. This issue may occur in the confidence-based uncertainty estimation methods, where a CNN is trained to predict aleatoric uncertainty

via binary classification. This tends to push the predicted scores to the extreme values zero and one and leads to large sets of pixels with the same uncertainty. The interested reader is referred to (Mehltretter and Heipke, 2021) for the interval-based AUC calculation.

5.4.2 Correlation Coefficient

The downside 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 simply neglects the actual magnitude of the estimated uncertainty and thus also the relation between the estimates and the real disparity error. To overcome this limitation, we therefore use the Pearson correlation coefficient between the absolute disparity error and the estimated uncertainty to evaluate the uncertainty models following Mehltretter and Heipke (2021). The Pearson correlation coefficient is a measure of linear dependence between two variables A and B and is defined as:

$$\rho(A,B) = \frac{\sum_{i=1}^{N}(A_i - \mu_A)(B_i - \mu_B)}{\sqrt{\sum_{i=1}^{N}(A_i - \mu_A)^2 \sum_{i=1}^{N}(B_i - \mu_B)^2}} \tag{5.2}$$

where μ and σ are the mean and standard deviation of the corresponding variable. This measure returns a value between -1 and 1, where 1 indicates a strong positive relationship, -1 indicates a strong negative relationship and 0 indicates no relationship at all. Thus, the higher the correlation, the better the uncertainty can be quantified.

5.4.3 n-MeRCI

The correlation coefficient metric enables the measure of linear dependence between the absolute disparity error and the uncertainty estimate. However, it is often recognized that outliers can occur in the dataset. The correlation coefficient ρ in Equation 5.2 is based on the sample means μ_A and μ_B, respectively, which are known to be very sensitive to the presence of outliers (Abdullah, 1990). To ensure the validity of the conclusions we draw in this work, the normalized mean rescaled confidence interval (n-MeRCI) (Moukari et al., 2019) is utilised as an additional metric to evaluate the correlation between the predictive uncertainty and the true error.

According to Moukari et al. (2019), to be able to compare methods with different ranges of uncertainties, all uncertainty values σ_i are rescaled by a common factor λ^τ so that the rescaled uncertainties $\tilde{\sigma}_i = \lambda^\tau \sigma_i$ are compatible with the observed errors $|d_i - \hat{d}_i| \leq \tilde{\sigma}_i$. To be robust to outliers, this metric is verified by only a certain number $\tau\%$ of the samples. The inliers are extracted according to the sorted ratios $\lambda_i = \frac{|d_i - \hat{d}_i|}{\sigma_i}$ and the factor λ^τ is then the biggest λ of the extracted

inliers. The formulation of the mean rescaled confidence interval is defined as

$$\text{MeRCI}^{\tau} = \frac{1}{N} \sum_{i=1}^{N} \lambda^{\tau} \sigma_i. \tag{5.3}$$

To give meaningful and easily interpretable values, the MeRCI score is normalised using

$$\text{n-MeRCI}^{\tau} = \frac{\text{MeRCI}^{\tau} - MAE}{\max^{\tau}(|d_i - \hat{d}_i|) - MAE}, \tag{5.4}$$

where MAE denotes the mean absolute error of disparity estimates and $\max^{\tau}(\cdot)$ is computed on the same extracted τ^{th} percentile, as for the MeRCI score. The resulting n-MeRCI score is equal to 1 when the network predicts a constant value for all data points, and is equal to 0 when the estimated uncertainty equal to the actual observed error. It is noteworthy that this score can be greater than 1. The original authors characterize the uncertainty predictor in this case as useless, since the score is worse than that of a constant uncertainty predictor.

5.5 Evaluation Procedure

To ensure a clear separation of training and test data, we evaluate the four uncertainty models, which are trained on the KITTI 2012 dataset, on the KITTI 2015 and Middlebury v3 datasets. According to (Mehltretter and Heipke, 2021), the depth of the input cost volume for the network is set to 192 pixels and the image resolution is halved during testing, as long as the maximum disparity exceeds 192. Thus, the cost volumes of the Middlebury dataset correspond to images with one quarter of the original resolution, whereas on the KITTI 2015 dataset, cost volumes correspond to images in the original resolution. Moreover, since the geometry-aware model assumes different uncertainty models for "good" and "hard" regions within a scene, respectively, we distinguish between these two types of regions in our complete evaluation to allow a fair comparison and a differentiated analysis of the results.

In this work, we consider the standard deviation of each assumed probability distribution as the uncertainty. As the log standard deviation s is predicted in the Laplacian, geometry-aware and mixture model, the standard deviation σ is then computed using $\sigma = \exp(s)$. Since the mixture model assumes a mixture of Laplace and Uniform distribution with the same mean, i.e. the estimated disparity d ($\mu_{\mathcal{L}} = \mu_{\mathcal{U}} = d$), and additionally predicts the inlier probability α, the standard deviation of the mixture model can be calculated using Equation 2.14. Thus, it becomes:

$$\begin{aligned} \sigma_{\text{Mixture}} = \sqrt{\sigma^2_{\text{Mixture}}} &= \sqrt{\alpha(\sigma^2_{\mathcal{L}} + \mu^2_{\mathcal{L}}) + (1-\alpha)(\sigma^2_{\mathcal{U}} + \mu^2_{\mathcal{U}})} \\ &= \sqrt{\alpha \cdot \exp^2(s_{\mathcal{L}}) + (1-\alpha)\exp^2(s_U)}. \end{aligned} \tag{5.5}$$

Instead of log standard deviation, the GMM model predicts the log variance s_k along with the mean μ_k and the mixture coefficient ϕ_k for each Gaussian component k. Thus, the variance of k-th Gaussian component is computed using $\sigma_k^2 = \exp(s_k)$. According to Equation 2.14, the mean μ_{GMM} and the standard deviation σ_{GMM} are defined as:

$$\mu_{\text{GMM}} = \sum_{k=1}^{K} \phi_k \, \mu_k \quad \text{and} \quad \sigma_{\text{GMM}} = \sqrt{\left(\sum_{k=1}^{K} \phi_k \left(\exp(s_k) + \mu_k^2 \right) \right) - \mu_{\text{GMM}}^2}. \tag{5.6}$$

6 Results

In this chapter, the three variants proposed earlier in this work (see Sec. 4.3) are first analysed along with the state-of-the-art Laplacian variant with respect to the AUC metric and correlation coefficient (see Sec. 5.5) in Section 6.1. In Section 6.2, we compare and analyse the different uncertainty models based on the normalised mean rescaled confidence interval (n-MeRCI) metric.

6.1 AUC and Correlation Coefficient

The different uncertainty models are first compared based on the AUC and correlation coefficient metrics in Section 6.1.1. Due to the large performance drop of the GMM model trained on cost volumes generated by MC-CNN, we investigate this phenomenon and describe our analysis in the subsequent Section 6.1.2.

6.1.1 Comparison of the Uncertainty Models

The AUC values of the four uncertainty models on the two datasets based on two disparity methods are shown in Table 6.1 along with the corresponding optimal values. As the geometry-aware model assumes different uncertainty distributions for "good" and "hard" regions, we distinguish between these two types of regions in the following evaluation.

As shown in Table 6.1, the AUC values of "hard" regions are always significantly larger than the values of "good" regions, indicating the higher error rate of "hard" regions, which can also be seen in the ROC curves in Fig. 6.1. This observation confirms that these "hard" regions are challenging for the disparity methods, which is the fundamental motivation for aleatoric uncertainty estimation. On the other hand, the differences among the Laplacian, the geometry-aware and the mixture uncertainty models with respect to the AUC values shown in Table 6.1 are not distinct. On all evaluated combinations of disparity methods and datasets the results of these three models are relatively similar. However, the AUC values of the GMM uncertainty model largely exceed those of the other three models. Analysing the ROC curves in Figure 6.1, it can be seen that while the curves of the other three models approach the optima, the GMM model already peaks at the first sampling. It indicates that the GMM model cannot distinguish the correct disparity estimates from

avg. AUC	Opt.		Laplacian		Geometry		GMM		Mixture	
$= 10^{-2}\times$	good	hard	good	hard	good	hard	good	hard	good	hard
KITTI 2015 Menze and Geiger (2015)										
Census-BM	5.69	22.13	7.94	28.05	**7.92**	**27.83**	18.04	45.00	8.10	27.96
MC-CNN	1.06	7.19	**2.19**	10.35	2.40	11.04	5.82	20.03	2.27	**10.27**
Middlebury v3 Scharstein et al. (2014)										
Census-BM	1.66	14.26	**3.38**	**18.70**	3.68	18.89	10.16	34.18	3.70	18.87
MC-CNN	0.51	8.94	1.51	12.03	1.44	**11.59**	3.36	17.60	**1.36**	12.09

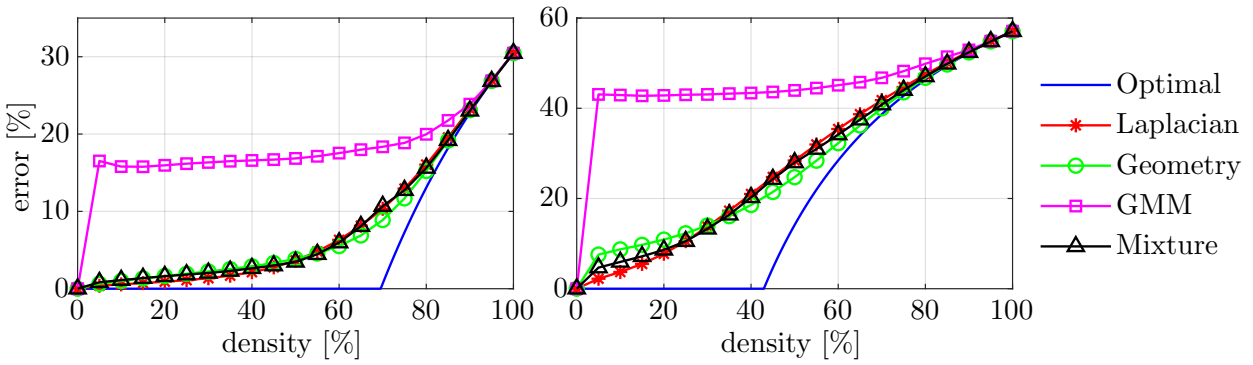
Table 6.1: Comparison of the four uncertainty models based on the AUC metric. The smaller the values, the better, while Opt. is the best achievable value (c.f. Sec. 5.4.1). These results are based on the first 100 images of the KITTI 2015 dataset and all images of the Middlebury v3 dataset, except for the configuration *MC-CNN + Middlebury v3*. Due to a noticeable domain gap for this configuration, we fine-tune the network on the first 4 images and test on the last 10 images of the Middlebury v3 dataset.

the incorrect ones properly. Moreover, it can also be seen that our geometry-aware model performs slightly better in detecting erroneous depth estimates in the region of medium density and a bit worse in the region of low density compared to the Laplacian baseline.

As already mentioned in Section 5.5, the AUC metric neglects the relation between the error magnitude and the predicted uncertainty by only considering the ability to separate correct from incorrect disparity assignments, which limits its expressiveness. Thus, we additionally investigate the relation between the absolute disparity error of a pixel and the corresponding predicted uncertainty using the correlation coefficient. The results on the two datasets as well as on the cost volumes generated by two disparity methods are shown in Table 6.2.

Analysing the correlation coefficients presented in Table 6.2, it can be seen that the geometry-aware and the mixture uncertainty models proposed in this work outperform the state-of-the-art Laplacian model in all configurations evaluated. Especially the geometry-aware model exceeds the Laplacian one by a wide margin, which can also be seen in Figure 6.2: While the heatmaps of the other three models are more dispersed for both "good" and "hard" regions, the results of the geometry-aware model shows significantly stronger correlations between the absolute disparity error and the predicted uncertainty. This supports our assumption regarding occluded and texture-less regions and demonstrates the benefit of additionally introducing a Uniform distribution to the loss term.

It is noteworthy that the correlation coefficients of our geometry-aware model are superior not only in the "hard" regions, but also in the "good" regions compared to the baseline, i.e. the Laplacian model. The reason for this phenomenon is that the Laplacian model tends to find a compromise



(a) "Good" region (b) "Hard" region

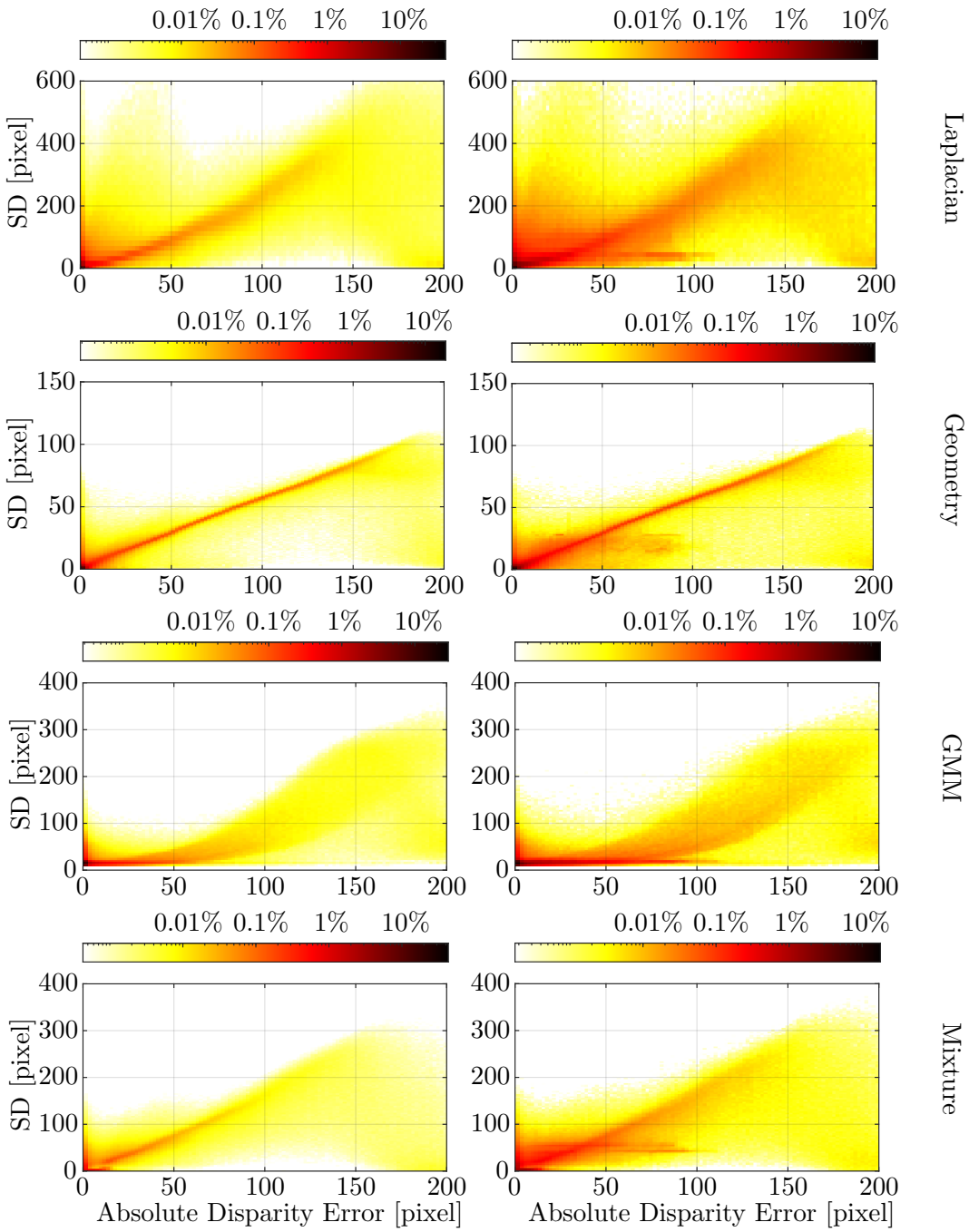
Figure 6.1: ROC curves for the first 100 images of the KITTI 2015 dataset and Census-based block matching. The closer the curve of an uncertainty model reaches the optimal curve, the higher the accuracy. The areas under these curves correspond to the AUC values in Table 6.1.

Correlation coefficient	Laplacian		Geometry		GMM		Mixture	
	good	hard	good	hard	good	hard	good	hard
KITTI 2015 Menze and Geiger (2015)								
Census-BM	0.74	0.72	**0.84**	**0.80**	0.81	0.77	0.78	0.77
MC-CNN	0.70	0.70	**0.73**	**0.73**	0.64	0.59	0.71	0.70
Middlebury v3 Scharstein et al. (2014)								
Census-BM	0.73	0.71	**0.82**	**0.81**	0.75	0.77	0.79	0.79
MC-CNN	0.68	0.71	**0.71**	**0.78**	0.60	0.68	0.70	0.74

Table 6.2: Comparison of the four uncertainty models based on the correlation coefficient, measuring the correlation between the absolute disparity error and the estimated uncertainty. The bigger the values, the better. For details on the evaluation procedure, please refer to Table 6.1.

between these two types of regions. Generally, pixels located in the "hard" regions have more potential to have large disparity errors. If the two different types of regions are not distinguished and a mini-batch consisting of samples with large and small errors is fed into the network, the Laplacian model would tend to fit the samples with large disparity errors in order to minimise the overall loss value, since the loss values (negative log likelihood) of such samples are larger than that of samples with small errors. Consequently, the accuracy of the Laplacian model would be reduced in both types of regions.

While the proposed geometry-aware and mixture model show superior results in all configurations in Table 6.2, the GMM model performs very well on the census-based cost volumes and shows a clear deterioration on the cost volumes originating from MC-CNN. This is probably due to the unstable training process and will be discussed in detail in the following Section 6.1.2. According to the heatmaps of the GMM model in Figure 6.2, the GMM model predicts rarely uncertainty



(a) "Good" region (b) "Hard" region

Figure 6.2: Absolute error uncertainty relation. The results are based on the first 100 images of the KITTI 2015 dataset using cost volumes computed with Census-based block matching. From top to bottom, the results of Laplacian, Geometry, GMM and Mixture models are shown. The logarithmic colour scale encodes the percentage of pixels showing the respective error and estimated uncertainty (standard deviation).

values below 10 pixels, resulting in a blank area near the horizontal axis. Regarding this problem, an activity regularization for the s-layer (c.f. Sec. 4.3.3) may be beneficial to prevent too large predicted log variance. Moreover, the heatmaps of the GMM model start with an almost horizontal line in the low error area and end with a dispersed positive relation between absolute error and uncertainty. Consequently, the resulting AUC values are very large due to the fact that the GMM model cannot differentiate the correct disparity estimates from the estimates with small errors, although the GMM model shows a high correlation in the middle and large error area, which indicates the limitation of the AUC metric.

Figure 6.2 further shows that the Laplacian model tends to assign small uncertainties to pixels with a large disparity error, especially visible for the "hard" regions. An example illustrating this case is shown in Figure 6.3: In the "hard" regions, the Laplacian model underestimates the uncertainty clearly for the non-textured areas on the wall especially visible in the Census-based uncertainty maps. The same behaviour can be observed for the occluded area behind the right computer (highlighted by a green arrow) in the uncertainty maps for both disparity methods. On the other hand, the Laplacian model also tends to assign large uncertainties to pixels with a disparity error between 0 and 50 pixels, which is not the case for the three models proposed in this work (see Fig. 6.2). An example of this behaviour can be seen in the Census-based uncertainty maps corresponding to hard regions shown in Figure 6.4: Compared to the other models and the error map, the Laplacian model generates relatively noisy uncertainty estimates in the areas of the door and back of the car containing some extremely high values.

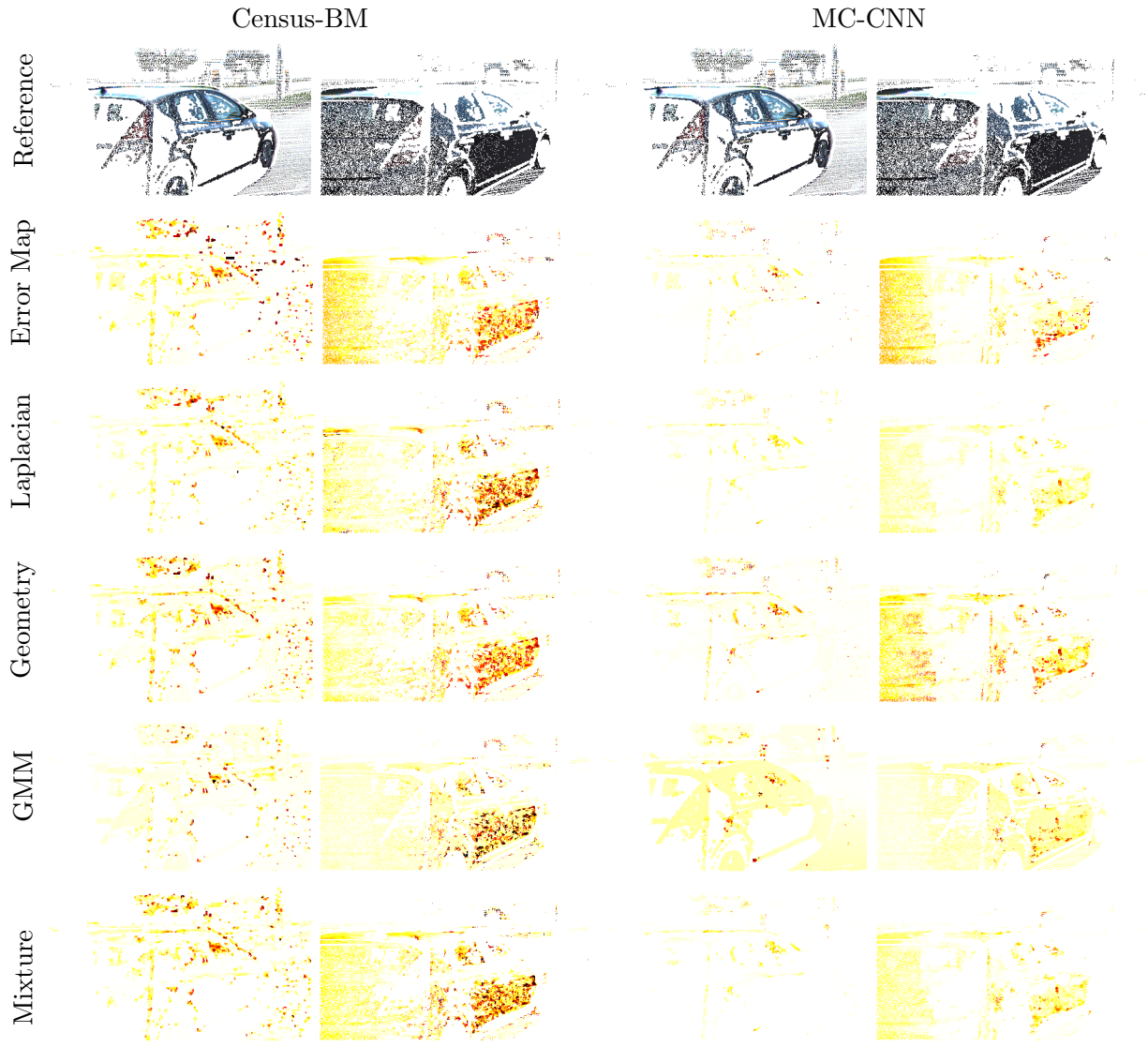
These two phenomenons can be explained by the fact that is described before: The Laplacian model tends to find a compromise between samples with large and small disparity errors. As pixels in the "hard" regions have more potential to have large disparity errors, these phenomenons indicate further the benefit of differentiation of the two types of regions. The differentiation can be addressed explicitly, e.g. by the geometry-aware model, or learned implicitly, e.g. by the GMM and the mixture model.

As the GMM model explicitly considers depth discontinuities, we also investigate the performance of all uncertainty models for pixels in such areas. Since the ground truth disparity maps of the KITTI 2015 dataset are not dense, we evaluate the performance for pixels at depth discontinuities of all uncertainty models on the Middlebury v3 dataset only. The masks for depth discontinuities are computed according to (Scharstein and Szeliski, 2002): pixels whose neighbouring disparities differ by more than 2 pixels, dilated by a 9×9 window. As show in Table 6.3, the results are similar to those in Table 6.2 that the proposed geometry-aware and mixture model demonstrate to be superior compared to the Laplacian baseline, while the GMM model has large performance drop on the cost volumes originating from the MC-CNN disparity method, probably due to the unstable training process (c.f. Sec. 6.1.2).

(a) "Good" region (b) "Hard" region (c) "Good" region (d) "Hard" region

Figure 6.3: Qualitative uncertainty evaluation on the Middlebury v3 dataset. The error and the uncertainty maps encode a high value in black and a small one in white. Note that also pixels without ground truth disparities are displayed in white. To allow a clear illustration of the spatial distribution of uncertainty, all uncertainty maps are normalised. In general, the proposed geometry model and mixture model show superior results in the "hard" regions, e.g. in the areas highlighted by green arrows, while uncertainties close to depth discontinuities (highlighted by red arrows) are less accurate for all four uncertainty models.

Although strong correlations between the absolute disparity error und the uncertainty are shown in Table 6.3, it can be seen that in some cases the uncertainties of pixels located at object boundaries are less accurate for all uncertainty models, especially can be seen in the uncertainty maps based on the MC-CNN disparity method in Figure 6.3 (area highlighted by a red arrow). This problem



(a) "Good" region (b) "Hard" region (c) "Good" region (d) "Hard" region

Figure 6.4: Qualitative uncertainty evaluation on the KITTI 2015 dataset. From left to right, the good and the hard regions based on Census-BM and MC-CNN disparity methods are shown. For details on the colour coding, please refer to Fig. 6.3. The uncertainty maps based on Census-BM generated by the three proposed models are smoother at the texture-less rear of the car, while the Laplacian model generates noisy uncertainties in this area.

is caused by the following potential reason: The assumed Laplacian or Laplace-Uniform mixture distribution is not suitable to describe the uncertainty related to depth discontinuities properly due to the multimodality that may occur in such areas. In this case, it was expected that a multimodal distribution, for example, a Gaussian mixture model, could better approximate the uncertainty at depth discontinuities. However, the GMM uncertainty model suffers from the unstable training process described in the following section and needs to be further investigated.

Correlation coefficient	Laplacian	Geometry	GMM	Mixture
Census-BM	0.71	**0.82**	0.77	0.79
MC-CNN	0.71	**0.77**	0.68	0.73

Table 6.3: Comparison of the four uncertainty models for pixels at depth discontinuities on the Middlebury v3 dataset using correlation coefficient. For details on the evaluation procedure, please refer to Table 6.1.

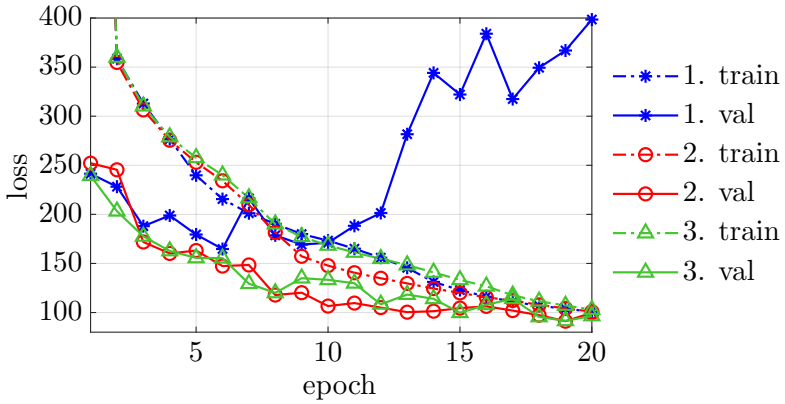


Figure 6.5: Comparison of the loss curves of three trained GMM uncertainty models based on Census-BM. While the train loss curves of all models converge to a similar loss value, the validation curve of the first model differs largely from other validation curves.

6.1.2 Investigation of the GMM Uncertainty Model

To investigate the large performance differences of the GMM model on the cost volumes of the two different disparity methods presented in Table 6.2 and 6.3, we retrain the GMM model twice based on Census-based block matching. The loss curves are shown in Figure 6.5 with the first GMM model corresponding to the GMM model that are used in Section 6.1.1. It can be seen in that an overfitting effect occurs very early in the first GMM model, i.e. the valuation loss increases while the train loss steadily decreases. However, this effect is not seen in other two models in the trained epochs. To further investigate the three GMM models, the correlation coefficients of these models are computed on the KITTI 2015 and Middlebury v3 dataset and are presented in Table 6.4.

As shown in Table 6.4, the first trained GMM model presents the best correlation between the absolute disparity error and the estimated uncertainty. It further shows that although the second and the third model have similar valuation curves (c.f. Fig. 6.5), their performance differ greatly. A possible reason for this phenomenon is that the Gaussian components of a assumed Gaussian mixture model are not fixed. Contrary to the distribution types assumed in other uncertainty models, whose mean value is strictly equal to the estimated disparity value, the positions of all Gaussian

Correlation coefficient	1. GMM		2. GMM		3. GMM	
	good	hard	good	hard	good	hard
KITTI 2015	**0.81**	**0.77**	0.77	0.74	0.64	0.61
Middlebury v3	**0.75**	**0.77**	0.57	0.72	0.59	0.57

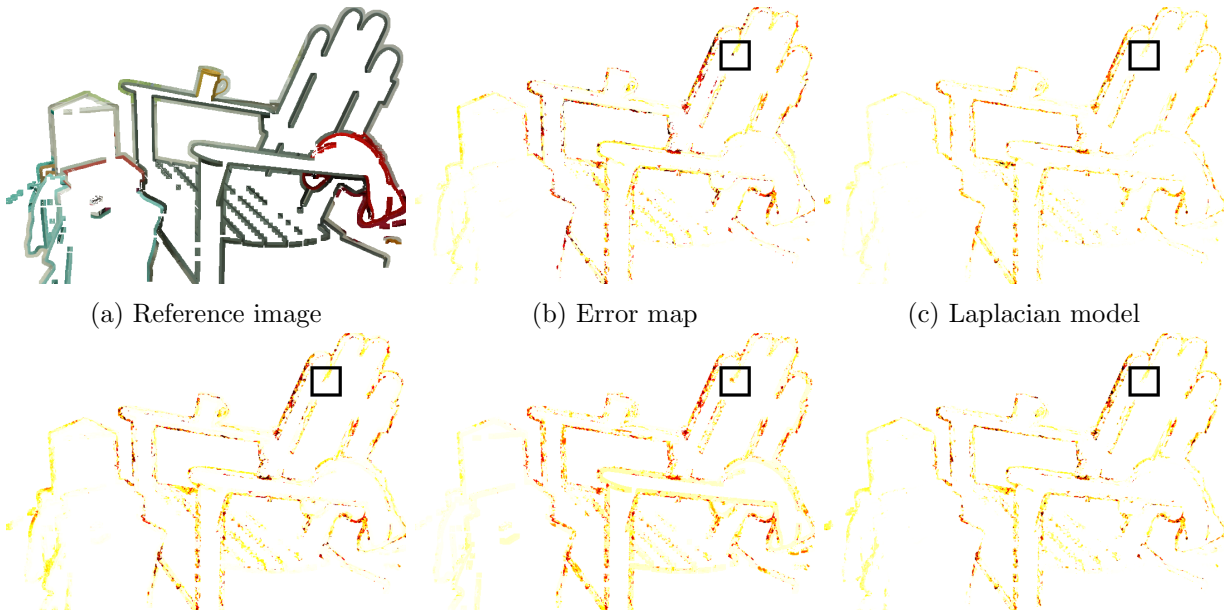
Table 6.4: Correlation coefficient comparison of the three trained GMM models using the same training setup based on the census-based cost volumes. These results are computed on the first 100 images of the KITTI 2015 dataset and all images of the Middlebury v3 dataset.

components in the GMM model can be permuted without changing the loss value. Consequently, the loss function may have many local minima, causing the convergence problem. Addressing this problem, a regularisation term can be added to the loss term to penalize the shift of the Gaussian components. Such a regularisation could force the first predicted mean to be the smallest mean, the second mean to be the second smallest, and so on.

Furthermore, the disparity maps previously computed based on the cost volumes using the "winner takes all" strategy, may constrain the GMM model. In other words, the GMM model could have learned the most important characteristics of a cost curve during training, however, we utilise the mean shift error (c.f. Sec. 4.3.3) to force the mean of the GMM model to approach the estimated disparity, which might reduce the learned characteristics and further reduce the accuracy. In this case, a network architecture predicting the disparity along with the uncertainty may be beneficial.

Despite the variations presented in Table 6.4 and the possible constraint, the first GMM model performs very well on the census-based cost volumes and shows a good generalization on both utilised datasets. On both datasets, the first GMM model presents comparable results to our proposed geometry-aware model, which is considered to be the best uncertainty model according to the correlation coefficients shown in Table 6.2. Moreover, a huge advantage of the GMM model against the geometry-aware model is that the outputs of the GMM model can be interpreted without the need of binary region masks. On the contrary, the geometry-aware model would need the information of which region the pixel corresponds to for a proper interpretation of the uncertainty of a pixel. Compared to the texture-less regions, whose masks are based on the RGB images, the computation of the masks for occluded regions needs the information regarding the ground truth disparities, which is unfortunately not applicable in the real-world applications.

Furthermore, the GMM model is the only model among the four uncertainty models within this work which can approximate the multimodality of a cost curve corresponding to a pixel in the left image of a stereo image pair. As the multimodality may occur in the region of depth discontinuities, we show such an example of the Middlebury v3 dataset in Figure 6.6. Note that the GMM model in this figure corresponds to the second model in Table 6.4, but the other two models show also similar results. As other uncertainty models misquantify the pixels marked by a black rectangle,



(a) Reference image (b) Error map (c) Laplacian model

(d) Geometry model (e) GMM model (f) Mixture model

Figure 6.6: Qualitative uncertainty evaluation for pixels at depth discontinuities on the Middlebury v3 dataset. These results are based on the Census-BM disparity method. For details on the colour coding, please refer to Fig. 6.3. While the uncertainty on the seat back (marked with a black rectangle) is misquantified by other uncertainty models, the GMM model shows a better match.

the predictions of the GMM model correspond to the real errors. Although the expressiveness of this behaviour is limited by the large variations of the GMM model caused by unstable training process, we can see the potential of the GMM model. Together with the superior performance of the first GMM model shown in Table 6.4 and the good generalization on both datasets, the GMM model is expected to have even better results once the unstable training problem is solved.

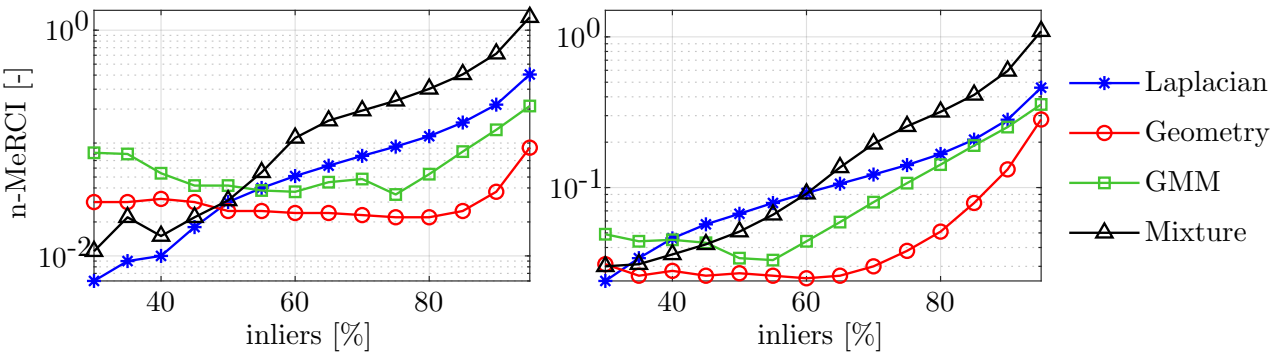
6.2 n-MeRCI

As the AUC and correlation coefficient metric are utilised to evaluate the estimated uncertainty in Section 6.1, in this section, we investigate the additional metric normalized mean rescaled confidence interval (n-MeRCI) (c.f. Sec. 5.4.3) for uncertainty evaluation. The results of the four uncertainty models on the two datasets as well as on the cost volumes generated by the two disparity methods are presented in Table 6.5.

As shown in Table 6.5, the proposed geometry-aware uncertainty model demonstrates the best quantification of aleatoric uncertainty values, which is similar to the results presented in Table 6.2. On the contrary to Table 6.2, the GMM model is considered to be the second best uncertainty model based on the n-MeRCI score, exceeding the baseline, i.e. the Laplacian model, in all configurations.

n-MeRCI	Laplacian		Geometry		GMM		Mixture	
	good	hard	good	hard	good	hard	good	hard
KITTI 2015 Menze and Geiger (2015)								
Census-BM	0.41	0.46	**0.09**	**0.28**	0.21	0.36	1.30	1.09
MC-CNN	0.18	0.20	**0.04**	**0.12**	0.06	0.19	0.38	0.51
Middlebury v3 Scharstein et al. (2014)								
Census-BM	0.17	0.53	**0.05**	0.31	0.10	**0.23**	0.45	1.16
MC-CNN	0.07	0.23	**0.04**	**0.17**	0.06	0.17	0.52	1.02

Table 6.5: Comparison of the four uncertainty models based on the n-MeRCI score assuming 95% of all uncertainty estimates are reliable. The smaller the values, the better. For details on the evaluation procedure, please refer to Table 6.1.



(a) "Good" region (b) "Hard" region

Figure 6.7: n-MeRCI score w.r.t. inlier rate based on the KITTI 2015 dataset and Census-BM disparity method. The n-MeRCI score is strongly dependent on the prior knowledge about the inlier rate of the uncertainty estimates, the disparity estimates as well as the ground truths.

The reason for this will be described in the following paragraphs. Furthermore, the mixture model shows a significant deterioration compared to Table 6.2 and has even scores larger than 1. Since a constant predictor would result in a n-MeRCI score of 1, Moukari et al. (2019) consider a uncertainty model with a n-MeRCI score larger than 1 as useless. However, the validity of the mixture model is already proven by the correlation coefficients as well as by the qualitative results shown in Figure 6.3 and 6.4. This raises our doubts about the validity of this metric. To investigate these phenomenons, we compute the n-MeRCI score in terms of the inlier rate τ (c.f. Sec. 5.4.3) and show the results based on the KITTI 2015 dataset and Census-BM disparity method in Figure 6.7.

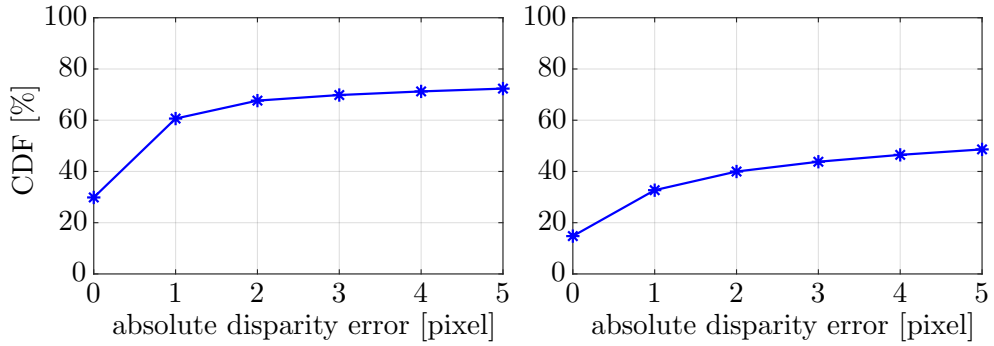
Figure 6.7 shows that the n-MeRCI score is strongly dependent on the assumed inlier rate. The statements could be totally different, if different inlier rates are set, for example, the statements would differ greatly before and after 50% in the "good" region. A proper inlier rate should be set for the n-MeRCI metric to allow a fair comparison. However, this requires the prior knowledge about

the inliers of the uncertainty estimates, which limits the expressiveness of this metric. Furthermore, it can be seen that the geometry-aware and the GMM model have turning points in both regions in Figure 6.7, for example, the point at 80% and at 75% for the geometry-aware and the GMM model in the "good" region, respectively, which are not seen in the Laplacian and the mixture model.

These turning points may indicate the true inlier rates for the respective model. However, they are also strictly dependent of the way to define an inlier. Moukari et al. (2019) use the division of absolute disparity error by uncertainty $\frac{|d_i - \hat{d}_i|}{\sigma_i}$ to define if an uncertainty estimate is reliable, which is very sensitive to small absolute disparity errors. Assuming two uncertainty values, for example, 0.001 and 0.1, are predicted by two uncertainty predictors for a pixel with an absolute disparity error of 1. The ratio $\frac{|d_i - \hat{d}_i|}{\sigma_i}$ of the first predictor is 100 times bigger than that of the second predictor, although both predicted uncertainty values are quite close to the true error. Since the n-MeRCI score is computed based on such ratios, the results could be strongly affected, especially, when many pixels with small absolute disparity errors occur in the dataset as illustrated in Figure 6.8.

Since the GMM model predicts rarely uncertainty values below 10 pixels (c.f. Figure 6.2), the ratio $\frac{|d_i - \hat{d}_i|}{\sigma_i}$ would be very small for a majority of pixels, whose absolute disparity errors are small, resulting in a smaller n-MeRCI score for the GMM model compared to other uncertainty models. Despite the sensitiveness of this metric to small absolute disparity errors, there is another possible reason for the deterioration of the mixture model shown in Table 6.5. The mixture model assumes a mixture of a Laplacian and a Uniform distribution. These two types of distributions have totally different interpretations of the standard deviation. Consequently, pixels with same absolute disparity errors may be interpreted with totally different uncertainty values, if different mixture coefficients are set, which makes it difficult to directly compare the mixture model with other uncertainty models. However, this phenomenon is not (or less) seen in the correlation-based experiments (c.f. Tab. 6.2 and Fig. 6.2) and needs further investigations.

Keeping the purposes of the n-MeRCI metric in mind, i.e. robustness and scale invariance (see Sec. 5.4.3), an approach which is less sensitive to pixels with small absolute disparity errors could be: 1) detecting the outliers using a metric combining the absolute difference between the uncertainty and the absolute disparity error with the magnitude of the absolute disparity error and 2) shifting the range of uncertainty values using normalisation, for example, min-max normalisation. However, this is beyond the scope of the present work and is subject of further investigations.



(a) "Good" region (b) "Hard" region

Figure 6.8: Cumulative density function (CDF) w.r.t. absolute disparity error in range [0, 5] based on the KITTI 2015 dataset and Census-BM disparity method. Pixels with small absolute disparity errors, e.g. $|d_i - \hat{d}_i| < 5$ pixels, account for a large proportion of all pixels, i.e. 72% and 49% for "good" and "hard" region, respectively.

7 Conclusion

In the present work, we propose three novel mixed probability models for the task of aleatoric uncertainty estimation in the context of dense stereo matching. For this purpose, we explicitly consider commonly challenging regions and outlier measurements employing Gaussian mixture model as well as mixtures of a Laplacian and a Uniform distribution. We argue that real-world scenarios typically violate the assumption of a unimodal distribution, commonly assumed by probabilistic uncertainty models in the literature. Thus, mixed probability models are better suited to describe the uncertainty inherent in such scenarios.

In our experiments, we use the architecture of CVA-Net, which utilises cost volume as input data, to investigate the effects of the proposed uncertainty models. We evaluate the performance of these models on two different datasets using cost volumes originating from two different dense stereo matching methods. To get as comprehensive a picture as possible of the properties of the uncertainty models, we investigate the results with three different metrics and point out the drawback of each metric. The results of the geometry-aware and the mixture model demonstrate to be superior compared to the unimodal baseline, which is especially visible for occluded and texture-less areas of an image. Although the proposed GMM model suffers from the unstable training problem, we show that the GMM model has much potential.

As the GMM model suffers from the unstable training problem, we expect better performance once the training process is optimized. Since the disparity maps estimated in advance could constrain the performance of the GMM model, it may be beneficial to learn the disparity and the uncertainty jointly. Furthermore, we show in this work that the additional evaluation metric *normalized mean rescaled confidence interval* (n-MeRCI) is sensitive to pixels with small absolute disparity errors. The statements based on this metric could be strongly affected especially when pixels with small errors account for a large proportion. A less sensitive approach would be beneficial to increase the validity of this metric. Despite the limitations mentioned above, the theoretical considerations could be confirmed by the experiments and a significant improvement of the uncertainty estimation could be achieved.

Bibliography

Abdullah, M. B., 1990. On a robust correlation coefficient. *Journal of the Royal Statistical Society* 39(4), pp. 1455–1460.

Achtelik, M., Bachrach, A., He, R., Prentice, S. and Roy, N., 2009. Stereo vision and laser odometry for autonomous helicopters in GPS-denied indoor environments. Vol. 7332, International Society for Optics and Photonics, SPIE, pp. 336–345.

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.

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg.

Carreira-Perpinan, M. A., 2000. Mode-finding for mixtures of gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(11), pp. 1318–1323.

Fu, Z., Ardabilian, M. and Stern, G., 2019. Stereo Matching Confidence Learning Based on Multimodal Convolution Neural Networks. In: *Representations, Analysis and Recognition of Shape and Motion from Imaging Data*, Springer, Cham, pp. 69–81.

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.

Girshick, R., 2015. Fast R-CNN. In: *IEEE International Conference on Computer Vision*, pp. 1440–1448.

Gneiting, T., Balabdaoui, F. and Raftery, A. E., 2007. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2), pp. 243–268.

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.

Huber, P. J., 1981. *Robust Statistics*. Wiley, New York.

Kendall, A. and Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: *Advances in Neural Information Processing Systems*, pp. 5574–5584.

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

Kuleshov, V., Fenner, N. and Ermon, S., 2018. Accurate Uncertainties for Deep Learning Using Calibrated Regression. In: *Proceedings of the International Conference on Machine Learning*, Vol. 80, pp. 2796–2804.

Lin, M., Chen, Q. and Yan, S., 2014. Network In Network. In: *Proceedings of the International Conference on Learning Representations*.

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.

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* pp. 63–75.

Mehltretter, M., Kleinschmidt, S. P., Wagner, B. and Heipke, C., 2018. Multimodal dense stereo matching. In: *German Conference on Pattern Recognition*, Springer, pp. 407–421.

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.

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: *International Conference on Intelligent Robots and Systems*.

Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S. and Fitzgibbon, A., 2011. KinectFusion: Real-time dense surface mapping and tracking. In: *IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136.

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.

Poggi, M., Aleotti, F., Tosi, F. and Mattoccia, S., 2020. On the Uncertainty of Self-Supervised Monocular Depth Estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3227–3237.

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

Poggi, M. and Mattoccia, S., 2016b. Learning from scratch a confidence measure. In: *Proceedings of the British Machine Vision Conference*, BMVA Press, pp. 46.1–46.13.

Scharstein, D. and Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* 47(1), pp. 7–42.

Scharstein, D., Hirschmueller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X. and Westling, P., 2014. High-resolution Stereo Datasets with Subpixel-accurate Ground Truth. In: *German Conference on Pattern Recognition*, Springer, pp. 31–42.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(56), pp. 1929–1958.

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.

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.

Vogiatzis, G. and Hernandez, C., 2011. Video-based, real-time multi-view stereo. *Image and Vision Computing*.

Zabih, R. and Woodfill, J., 1994. Non-Parametric Local Transforms for Computing Visual Correspondence. In: *Proceedings of the European Conference on Computer Vision*, Springer, pp. 151–158.

Zbontar, J., LeCun, Y. et al., 2016. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research* 17(1), pp. 2287–2318.