JOINT ESTIMATION OF DEPTH AND ITS UNCERTAINTY FROM STEREO IMAGES USING BAYESIAN DEEP LEARNING

Max Mehltretter

Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Germany
mehltretter@ipi.uni-hannover.de

Commission II, WG II/2

KEY WORDS: Dense Matching, Depth Reconstruction, Uncertainty Quantification, Deep Learning, Bayesian Neural Network

ABSTRACT:

The necessity to identify errors in the context of image-based 3D reconstruction has motivated the development of various methods for the estimation of uncertainty associated with depth estimates in recent years. Most of these methods exclusively estimate aleatoric uncertainty, which describes stochastic effects. On the other hand, epistemic uncertainty, which accounts for simplifications or incorrect assumptions with respect to the formulated model hypothesis, is often neglected. However, to accurately quantify the uncertainty inherent in a process, it is necessary to consider all potential sources of uncertainty and to model their stochastic behaviour appropriately. To approach this objective, a holistic method to jointly estimate disparity and uncertainty is presented in this work, taking into account both aleatoric and epistemic uncertainty. For this purpose, the proposed method is based on a Bayesian Neural Network, which is trained with variational inference using a probabilistic loss formulation. To evaluate the performance of the method proposed, extensive experiments are carried out on three datasets considering real-world indoor and outdoor scenes. The results of these experiments demonstrate that the proposed method is able to estimate the uncertainty accurately, while showing a similar and for some scenarios improved depth estimation capability compared to the dense stereo matching approach used as deterministic baseline. Moreover, the evaluation reveals the importance of considering both, aleatoric and epistemic uncertainty, in order to achieve an accurate estimation of the overall uncertainty related to a depth estimate.

1. INTRODUCTION

Reconstructing the 3D geometry of a scene from stereo images is a fundamental task in photogrammetry and computer vision, commonly forming the basis for higher-level tasks that build on the estimated depth information. While current dense stereo matching methods demonstrate convincing results, with deep learning-based approaches showing a particularly low number of erroneous estimates, the resulting depth information is not free of errors. Especially image regions with a weak texture, that are occluded or that are located close to depth discontinuities remain challenging and may cause errors in the dense stereo matching procedure. In order to not propagate such errors to higher-level tasks unknowingly, a measure of the uncertainty associated to a depth estimate is needed. In turn, tasks that rely on image-based depth information, such as 3D pedestrian tracking (Nguyen and Heipke, 2020) or the estimation of the pose and shape of a vehicle (Coenen and Rottensteiner, 2021), may be improved if the associated uncertainty is known. On the other hand, knowledge on the uncertainty might even be a crucial prerequisite for safety-critical tasks, for example, for applications from the domain of autonomous driving.

Following the taxonomy proposed by Hacking (1975), two types of uncertainties can, in general, be distinguished: aleatoric and epistemic uncertainty. Aleatoric uncertainty is contained in the data and caused by variable, non-deterministic or simply unpredictable behaviour of a process under consideration. In contrast, epistemic uncertainty accounts for incorrect or inaccurate model hypotheses. From the perspective of dense stereo matching, aleatoric uncertainty accounts for effects such as sensor noise, occlusion and matching ambiguities caused, for example, by texture-less areas or repetitive patterns within a scene. Epistemic uncertainty, on the other hand, considers assumptions that simplify the matching process and characteristics that are missing in the definition of this process (or in case of deep learning-based approaches in the training data), such as features and shades that imply a certain geometric shape. Note that the assignment of specific effects to either aleatoric or epistemic uncertainty is not fixed, but depends on the definition of the problem domain. In the literature, several methods have been presented addressing aleatoric uncertainty estimation for dense stereo matching. These methods cover a wide range of functional and stochastic models, while commonly representing aleatoric uncertainty either as confidence, a unit-less measure of reliability in the range between zero and one, or as standard deviation of a particular probability distribution. In contrast, the estimation of epistemic uncertainty, and thus also the joint estimation of aleatoric and epistemic uncertainty, is much rarely discussed in the literature with respect to more complex photogrammetric tasks, which is particularly true for dense stereo matching. However, to accurately estimate the uncertainty embedded in a process, it is necessary to consider all potential sources of uncertainty.

To overcome this limitation, a holistic method to jointly estimate depth and uncertainty is presented in this work, considering aleatoric and epistemic uncertainty, both modelled in a Bayesian deep learning framework. Thus, the main contribution of this work is the realisation of a Bayesian Neural Network (BNN), in terms of the definition of a functional and a stochastic model. The functional model is characterised by probabilistic convolutional layers that are trained using Variational Inference (VI) and that are incorporated into an end-to-end trainable Convolutional Neural Network (CNN) architecture, which has already proven to be well-suited for the task

of dense stereo matching. The stochastic model is based on a naive mean field approximation, assuming a variational distribution that consists of an independent Gaussian distribution for each parameter of the probabilistic convolutional layers. The loss function proposed is formulated in a way that the estimation of all three values, depth, aleatoric and epistemic uncertainty, can be learned jointly and end-to-end. For this purpose, the ideas of likelihood maximisation under a specific mixture distribution and the minimisation of the Kullback-Leibler (KL) divergence between the assumed variational distribution and the exact posterior are combined.

2. RELATED WORK

In recent years, many publications have been presented in the literature addressing uncertainty estimation in the context of dense stereo matching. While this emphasises the relevance and the actuality of this topic, most of these works focus exclusively on the estimation of aleatoric uncertainty. As shown by Hu and Mordohai (2012) and Poggi et al. (2021) in comprehensive evaluations, a multitude of approaches has been proposed for this task, using the stereo images directly or various (intermediate) representations of the dense stereo matching procedure, such as a cost volume or a disparity map, as input. The aleatoric uncertainty itself is typically represented either as confidence, a value between zero and one that represents how trustworthy the associated depth estimate is, or as the standard deviation of a particular probability distribution. While most works in the literature measure aleatoric uncertainty in terms of confidence (Hu and Mordohai, 2012; Kim et al., 2019), this approach does not allow to assess the uncertainty in pixels or metric units and thus prevents to reason about the actual error magnitude. On the other hand, aleatoric uncertainty can be learned in a Bayesian way via maximum likelihood estimation (Kendall and Gal, 2017). Following this approach, the parameters of a particular probability distribution, for example, mean and standard deviation of a Gaussian distribution over the disparity, are understood as predictions, while maximising the likelihood of the ground truth disparity during training. Different types of distributions have been proposed for this purpose, with an approach based on a mixture of a Laplace and a Uniform distribution, which considers the geometry and appearance of the observed scene, showing the best results so far (Zhong and Mehltretter, 2021). However, almost all of these methods model aleatoric uncertainty estimation as a separate task that is carried out subsequent to the actual dense stereo matching. While this procedure simplifies the individual tasks, as only one value has to be estimated at a time while the other one is kept constant, it prevents from exploiting synergies that may arise from the joint estimation of depth and the associated aleatoric uncertainty.

Compared to aleatoric uncertainty, which is commonly treated as an additional predictive value, the estimation of epistemic uncertainty is typically more difficult and is addressed far less frequently in the literature. However, this type of uncertainty helps to mitigate the problem of overconfident predictions and to identify cases in which a method is highly uncertain regarding its prediction, for example, processing data outside of the learned data distribution. To cope with this task, in particular the use of stochastic neural networks has proven to be well suited, which allow to learn a distribution over the parameters, instead of learning point estimates as parameter values (Jospin et al., 2020). Epistemic uncertainty is then commonly estimated via Monte Carlo sampling by deriving the central moments of the probability distribution describing the final result from the aggregation of the individual samples. Common realisations of stochastic neural networks are ensembles of deterministic neural networks that have been trained independently, Monte Carlo dropout and BNNs. Ensemble learning is the simplest of these three concepts, using varying seed values (Lakshminarayanan et al., 2017), different subsets of the training data (Moukari et al., 2019) or the parameter values of the same network obtained after various numbers of training epochs as individual networks to form an ensemble (Huang et al., 2017). On the other hand, Monte Carlo dropout, as used in (Kendall and Gal, 2017), is similar to classical dropout used for the purpose of regularisation during training, but applies this procedure not only during training but also at test time. Placing a Bernoulli distribution over the network weights, the weights are set to zero with a certain probability, resulting in a slightly different parametrisation of the same network for every forward pass.

BNNs constitute the third and last realisation of stochastic neural networks being discussed in this section that allows to define a prior for the parameters of the network, treating the uncertainty in a Bayesian manner. Despite the fact that the basic concepts of BNNs are already known for decades (MacKay, 1992), they have only recently been used in practice for more complex tasks, such as image-based object classification (Brosse et al., 2020). While ensemble learning as well as Monte Carlo dropout generally have the limitation that prior knowledge and the correlation between parameters of the network can not be considered, both is in principle possible using a BNN. A first step of using a BNN in the context of dense stereo matching was taken by Mehltretter (2020): They describe a BNN-based approach that allows to jointly estimate depth and aleatoric as well as epistemic uncertainty, characterising it as being close to the approach followed in the present work. Despite the good results for the epistemic uncertainty estimates, the authors state that the joint estimation of depth and aleatoric uncertainty leads to a deterioration of the depth estimation capability. This limitation is the main motivation for the present work and is aimed to be overcome by the methodology presented.

3. METHODOLOGY

In this section, a novel method to estimate depth and its associated aleatoric and epistemic uncertainty in the context of dense stereo matching is proposed, which is based on Bayesian deep learning. The input to the proposed method are stereoscopic image pairs $(\mathbf{I}_L, \mathbf{I}_R)$, referring to the left image $\mathbf{I}_L$ of such a pair as the reference image. It is assumed that both images were captured simultaneously, allowing to neglect the influence of movements of parts of the scene depicted, and have a reasonable overlap in which the depth can be determined via triangulation. Moreover, the stereo image pairs are presented to the proposed method after planar rectification, assuming that the interior orientations of both cameras and the relative orientation between them is known. In the following, the functional model in form of a BNN architecture is introduced first, before the stochastic model is described.

3.1 Functional Model

The functional model of the method presented in this work is defined as a BNN and is based on two CNN architectures presented in the literature: Geometry and Context Network (GC-Net) proposed by Kendall et al. (2017) and Cost Volume Analysis Network (CVA-Net) proposed by Mehltretter and

Heipke (2021). GC-Net is a dense stereo matching approach that follows the classical taxonomy of Scharstein and Szeliski (2002): First, features are extracted from the left and right image using a Siamese architecture consisting of multiple 2D convolutional layers with residual connections. In the second step, a cost volume is built by concatenating a feature vector from the left image with a feature vector from the right image for all potential point correspondences, defined by the corresponding horizontal epipolar line and the specified disparity range. This initial cost volume is further processed using 3D convolutional and transposed convolutional layers arranged in an encoder-decoder structure with skip connections. The output of this structure is a 3D cost volume similar to the one computed by conventional dense stereo matching approaches, from which a disparity map is extracted using a differentiable soft argmin layer. On the other hand, CVA-Net allows to estimate aleatoric uncertainty associated to depth estimates based on a cost volume (extract). For this purpose, several layers of 3D convolutions are used to initially combine cost information from a spatial local neighbourhood, before processing the result of this combination along the depth axis per pixel. To obtain an aleatoric uncertainty estimate per pixel, global average pooling is applied which reduces the processed 3D cost volume into a 2D uncertainty map. Both architectures are chosen because of their good accuracy for the tasks of dense stereo matching and aleatoric uncertainty estimation, respectively, while having a relatively low number of parameters.

The fusion of the two CNN architectures described is realised by adding CVA-Net as aleatoric uncertainty estimation branch to GC-Net, which runs in parallel to the soft argmin layer (see Fig. 1). While the basic structures of both architectures remain unchanged, CVA-Net receives the whole optimised cost volume, instead of operating on a cost volume extract as originally proposed by Mehltretter and Heipke (2021), which is possible due to the fully convolutional character of this CNN architecture. To transform the combined architecture from a CNN into a BNN, the parameters of the network are no longer learned directly, as it is done by conventional deep learning and which would result in constant point estimates for every parameter, but sampled from a probability distribution which is defined by the stochastic model presented in the following section. In this context, the network parameters θ are sampled anew for every individual forward pass k, which results in slightly different variants of the same network f_θ and thus in disparity maps $\mathbf{D}$ and aleatoric uncertainty maps $\mathbf{U}_A$ that vary with each sample:

$$f_{\theta_k}(\mathbf{I}_L, \mathbf{I}_R) = (\mathbf{D}_k, \mathbf{U}_{A,k})\,. \tag{1}$$

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

$$\mathbf{D}(\mathbf{p}) = \bar{d}_\mathbf{p} = \frac{1}{K}\sum_{k=1}^{K} d_{\mathbf{p},k}\,, \tag{2}$$

$$\mathbf{U}_E(\mathbf{p}) = \sigma^2_{E,\mathbf{p}} = \frac{1}{K-1}\sum_{k=1}^{K}(d_{\mathbf{p},k} - \bar{d}_\mathbf{p})^2\,. \tag{3}$$

Aggregating the resulting disparity estimates d of a pixel $\mathbf{p}$ over k samples, the average disparity estimate $\bar{d}$ and the variance σ^2_E are used to obtain a disparity map $\mathbf{D}$ and an epistemic uncertainty map $\mathbf{U}_E$, respectively. This procedure is justified by the observation that deviations between different disparity estimates assigned to the same pixel reflect the model's uncertainty to determine the correct disparity, which allows to approximate the epistemic uncertainty based on these deviations. Because the aleatoric uncertainty estimates vary with each Monte Carlo sample as well, it is necessary to aggregate the aleatoric uncertainty maps of all samples drawn to obtain a consistent result:

$$\mathbf{U}_A(\mathbf{p}) = \frac{1}{K}\sum_{k=1}^{K}\sigma^2_{A,\mathbf{p},k}\,, \tag{4}$$

where σ^2_A represents the aleatoric uncertainty computed according to the probabilistic model described in the next section.

Similar to the concepts presented in (Brosse et al., 2020; Mehltretter, 2020), not all parameters of the network architecture presented are treated in a probabilistic manner. While Brosse et al. (2020) argue that it is sufficient to only model the final layer(s) of an architecture probabilistically to assess the epistemic uncertainty and to benefit from the positive effect of ensemble learning on the accuracy, they only investigate such a setup in the context of classification. Preliminary experiments carried out in the context of this work and the results of (Mehltretter, 2020) have shown that a different approach is preferable for dense stereo matching: Only the weights belonging to convolutional filter kernels used in the feature extraction step (2D convolutions) and the multi-scale feature matching step in the encoder of the cost volume optimisation (3D convolutions) are treated probabilistically. In contrast, the parameters belonging to operations used to up-sample the intermediate feature maps (3D transposed convolutions), which is carried out in the decoder part of the cost volume optimisation step, are retained deterministically (cf. Fig. 1). Compared to treating all parameters in a probabilistic manner, the proposed procedure reduces the number of trainable parameters and the computational effort.

Besides the desired capability to estimate epistemic uncertainty, treating some parts of the network in a probabilistic manner further allows to reduce the model capacity without decreasing the accuracy of the estimated disparity maps. For this purpose, the number of filter channels n_c is adjusted, which is set to $n_c = 32$ for almost all layers of the original GC-Net architecture and to multiples of 32 if the spatial resolution of the feature maps is reduced in the inner layers of the encoder-decoder structure. As shown by the results of preliminary experiments, n_c can be reduced by 25% to 24 channels without affecting the performance of the described probabilistic variant, while this adjustment decreases the accuracy of the deterministic baseline. Such an adaptation of n_c reduces the number of parameters of the network as well as the size of the intermediate feature maps and thus the memory footprint and the computational effort. In summary, the proposed transformation of the described combination of GC-Net and CVA-Net into a probabilistic variant using 24 filter channels increases the number of parameters to be learned only marginally from about 3.6 to 3.7 Mio. (assuming that the stochastic model is defined as described in the next section).

3.2 Stochastic Model

To use the previously defined BNN for the purpose of Bayesian inference, the posterior distribution $p(\theta|\mathcal{D})$ of the network parameters θ given a set of training data $\mathcal{D}$ is required. However,

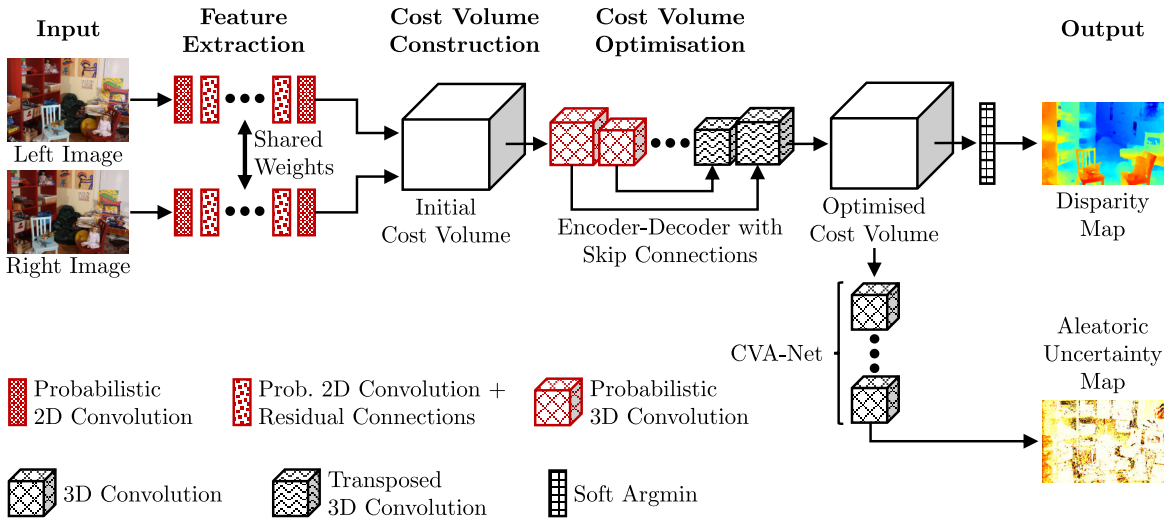


Figure 1. **Overview of the functional model.** While the probabilistic adaptation of the GC-Net architecture is trained to predict a disparity map corresponding to the left image of a planar rectified stereo image pair, the probabilistic convolutional layers further allow to estimate the corresponding epistemic uncertainty via Monte Carlo sampling. CVA-Net is integrated as a separate branch, operating on the optimised cost volume to additionally predict an aleatoric uncertainty map. Source: Adapted from Mehltretter (2020).

computing and thus also sampling from this exact posterior distribution is typically an intractable problem, due to the integral involved in the evidence which in general cannot be solved analytically. Therefore, variational inference is applied in this work, aiming to learn the parameters ϕ of a variational distribution q that approximates the exact posterior distribution. To measure the distance between the exact posterior distribution and its approximation, the KL divergence proposed by Kullback and Leibler (1951) is used, which is minimised during training in order to maximise the similarity of the two distributions.

To reduce the number of parameters to be learned and the computational overhead arising from VI compared to conventional deep learning, it is assume that the variational distribution over the latent variables, i.e., the network parameters, factorises as:

$$q(\theta_1, \theta_2, ..., \theta_n) = \prod_{i=1}^{n} q(\theta_i)\,. \tag{5}$$

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

$$\mathcal{L}_{\mathrm{KL}} = \sum_{i=1}^{n} KL(q_\phi(\theta_i)||p(\theta|\mathcal{D}))\,. \tag{6}$$

To further enable the proposed BNN to estimate aleatoric uncertainty, in addition to the disparity of a pixel and its associated epistemic uncertainty, the geometry-aware model of Zhong and Mehltretter (2021) is adapted in this work. This approach relies on the assumption that the aleatoric uncertainty associated to a pixel's disparity estimate can be represented by a probability distribution, which is characterised by a set of parameters that are predicted by a CNN. During training, the ability to predict these parameters is optimised with the objective of maximising the likelihood of the corresponding ground truth disparity under the probability distribution assumed (Kendall and Gal, 2017). Following this procedure, aleatoric uncertainty can be learned as standard deviation from the distribution of the disparity error, thus avoiding the need for a direct reference for the uncertainty, such as explicit parametrisations of the probability distribution. Note that contrary to the procedure proposed in the original publication of Zhong and Mehltretter (2021), the disparity estimate is not fixed in this work, but is optimised together with the aleatoric uncertainty.

The geometry-aware model of Zhong and Mehltretter (2021) is chosen due to its superior performance compared to other probabilistic models and its ability to adapt to challenging scenarios that are common in the context of dense stereo matching, such as occlusions and weakly textured areas. According to this model, the error of pixels that are expected to have a unique correspondence in the second image of a stereo pair (referred to as unique matching assumption) is assumed to be Laplacian distributed. On the other hand, errors arising from weakly textured and occluded regions are assumed to be uniformly distributed in specific intervals. Formulating these two assumptions as log likelihood terms, the following equations are obtained:

$$\mathcal{L}_{\mathrm{L}} = \frac{\sqrt{2}}{\exp(s_{\mathbf{p}})} |d_{\mathbf{p}} - \hat{d_{\mathbf{p}}}| + s_{\mathbf{p}}\,, \tag{7}$$

$$\mathcal{L}_{\mathrm{U}} = \begin{cases} 0.5x^2 & \text{if } |x| \leq \gamma \\ \gamma|x| - 0.5\gamma^2 & \text{otherwise}\,, \end{cases} \tag{8}$$

where d is the estimated and $\hat{d}$ the ground truth disparity, while s is the logarithm of the standard deviation of the assumed Laplace distribution. x is defined as the difference between the absolute disparity error and half the length of the interval with uniform distribution $r_{\mathbf{p}}$, resulting in: $x = |d_{\mathbf{p}} - \hat{d_{\mathbf{p}}}| - r_{\mathbf{p}}$.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2022-69-2022 | © Author(s) 2022. CC BY 4.0 License.

While the ground truth disparity $\hat{d}$ needs to lie in the interval $[d-r, d+r]$ to maximise the probability, x is minimised to prevent the network from predicting unreasonable large intervals. With the relationship between the interval length and the standard deviation σ_U of the uniform distribution, it further is: $r = \sqrt{3}\,\sigma_U$. The complete term $\mathcal{L}_{\mathrm{U}}$ is set up in form of a Huber loss function (Huber, 1981). Combining the two assumptions on the distribution of the disparity error, the following loss function can be obtained, which allows to train the proposed BNN end-to-end in a supervised manner using training data $\mathcal{D}$:

$$\mathcal{L}_{\text{Aleatoric}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{p} \in \mathcal{D}} \beta_{error} \cdot (c_i \cdot \mathcal{L}_{\mathrm{L}} + (1 - c_{\mathbf{p}}) \cdot \mathcal{L}_{\mathrm{U}} + \mathcal{L}_{\mathrm{BCE}}) \,, \quad (9)$$

$$\mathcal{L}_{\mathrm{BCE}} = \beta_{\mathbf{p}} \cdot h(o_{\mathbf{p}}, \hat{o}_{\mathbf{p}}) \,, \quad (10)$$

where c is a binary variable indicating whether the unique matching assumption is met or not. According to the definition of this binary classification discussed earlier, c is defined as: $c = \neg o \wedge \neg t$, where o specifies if the correspondence in the second image is occluded and t whether the pixel in the reference image is located in a weakly textured area. While t is determined based on the reference image directly using the criterion specified by Scharstein and Szeliski (2002), o is predicted by the CVA-Net branch of the proposed network in addition to the log standard deviation. In order to optimise the capability of predicting whether a pixel's correspondence is occluded or not, the loss function is extended by a binary cross-entropy term h, minimising the difference between the predicted and the reference occlusion values o and $\hat{o}$. In this context, the pixel-dependent weight $\beta_{\mathbf{p}}$ is defined as:

$$\beta_{\mathbf{p}} = \beta_{\mathrm{BCE}} \cdot (\hat{o}_{\mathbf{p}} \cdot (\beta_{\text{occluded}} - 1) + 1) \,. \quad (11)$$

It considers the class imbalance between non-occluded and occluded pixels using the ratio of their frequency in the training set as β_{occluded} as well as a static weight β_{BCE}, which is used to balance the influences of the binary cross-entropy term and the likelihood term. Compared to the original loss formulation by Zhong and Mehltretter (2021), we add a coefficient β_{error} used to weight the individual training samples according to their disparity error. This procedure is necessary if the error of the predicted disparities is not well distributed over the disparity range considered, but mainly concentrated around zero. While this is a desired behaviour in the context of dense stereo matching, it motivates to preferably predict small aleatoric uncertainties, thus resulting in an effect comparable to the one arising from imbalanced training samples in a classification setup.

Combining the different parts of the stochastic model that are necessary to estimate epistemic and aleatoric uncertainty and that have been described before, the following final loss formulation is obtained:

$$\mathcal{L}_{\text{Full-Uncertainty}} = \mathcal{L}_{\text{Aleatoric}} + \beta_{\mathrm{KL}} \cdot \mathcal{L}_{\mathrm{KL}} \,, \quad (12)$$

where β_{KL} is a hyper-parameter used to balance the two parts of the loss function. Relying on the concept of stochastic variational inference (Hoffman et al., 2013), the training procedure of the proposed BNN does not differ from the one used for ordinary CNNs in the sense that common optimisation algorithms can be applied. To mitigate the negative impact of stochastic sampling of parameters during training on the convergence behaviour, we apply Flipout as proposed by Wen et al. (2018).

Under the assumption that the aleatoric and the epistemic uncertainty are randomly and independently distributed, quadratic error propagation is applied to obtain the overall uncertainty associated with the disparity estimate of a pixel $\mathbf{p}$:

$$\sigma_{\mathbf{p}}^2 = \sigma_{A,\mathbf{p}}^2 + \sigma_{E,\mathbf{p}}^2 \,, \quad (13)$$

from which the definition of the overall uncertainty map $\mathbf{U}$ follows as $\mathbf{U} = \mathbf{U}_A + \mathbf{U}_E$. Consequently, following the method proposed, the estimation of disparity and aleatoric uncertainty is learned together exploiting the principle of likelihood maximisation, while the estimation of epistemic uncertainty is further enabled by the usage of a BNN trained via VI.

4. EXPERIMENTAL SETUP

In this section, the experimental setup used to evaluate the proposed methodology is described. For this purpose, the datasets used for training and testing are presented in Section 4.1. In Section 4.2, the framework for training the proposed approach is discussed, including an overview of the hyper-parameter settings. This section closes with a presentation of the strategy and criteria for testing in Section 4.3.

4.1 Datasets

In the experiments carried out in the context of this work, four different datasets have been used: Sceneflow FlyingThings3D (Mayer et al., 2016), InStereo2K (Bao et al., 2020), Middlebury stereo benchmark version 3 (Scharstein et al., 2014) and KITTI, which we define as the combination of the KITTI 2012 and 2015 stereo datasets (Geiger et al., 2012; Menze and Geiger, 2015). All these datasets consist of stereo image pairs with ground truth disparity maps corresponding to the reference image of each pair. The Sceneflow dataset contains about 27 thousand synthetic stereo image pairs that show abstract scenes with randomly located objects and provides a reference for the disparity for all pixels. The InStereo2K and Middlebury datasets consist of 2050 and 15 stereo image pairs, respectively, that show different indoor scenes. For both datasets, the reference for the disparity is captured via structured light and is provided for about 90% of the pixels. Lastly, the KITTI dataset consists of 394 stereo image pairs that show various street scenes and provides a reference for the disparity for about 30% of the pixels, which is derived from LIDAR point clouds.

4.2 Training Procedure

The BNN presented in this work is trained end-to-end in a fully supervised manner. Because of the large amount of training data necessary, the network is first trained for 24 epochs on 21 thousand synthetic stereo image pairs from the Sceneflow dataset, before it is fine-tuned for 57 epochs on 1800 real-world image pairs of the InStereo2K dataset. In each epoch, a random crop of size 384×96 pixels from every image pair is fed to the network using a mini-batch size of one. The optimum number of training epochs is determined via early stopping, i.e., the training procedure is terminated if the validation loss does not decrease in three consecutive epochs and the set of parameters associated to the epoch with minimum validation loss is used for testing. For both, training and fine-tuning, 100 images of the respective dataset are used as validation set. Moreover, in all training epochs and in the first 47 epochs of fine-tuning, the network is only optimised for the task of disparity estimation, neglecting the aleatoric uncertainty. For this purpose, the term $\mathcal{L}_{\text{Aleatoric}}$ in Equation 12 is replaced by the L1 loss. Only the last

10 epochs of fine-tuning are carried out using the loss function shown in Equation 12. This approach improves the convergence behaviour compared to directly optimising for both, disparity and aleatoric uncertainty, and leads to overall better results. The optimisation itself is realised using RMSProb (Tieleman and Hinton, 2012) with a learning rate of 10^{-3}.

The disparity range considered during training is limited to [0,191] pixels, thus pixels with a ground truth disparity outside of this range are discarded and not used for training the network parameters. The ratio between occluded and non-occluded pixels β_{occluded} used in Equation 11 is determined based on the ground truth disparity maps used for training and is set to $\beta_{\text{occluded}} = 20$. The parameter γ, which governs the transition between the two parts of the Huber loss in Equation 8, and the coefficient β_{BCE}, which weights the binary-cross entropy term relative to the likelihood term in Equation 11, are set to one. The Gaussian distributions that form the variational distribution and from which the parameters of the probabilistic 2D and 3D convolutional layers are sampled as $\theta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, are initialised with $\mu = 0$ and $\sigma^2 = 1$. In addition, all deterministic convolutional and transposed convolutional layers are initialised using the Glorot normal initialiser (Glorot and Bengio, 2010). The hyper-parameter β_{KL}, which is used to weight the KL divergence relative to the loss term $\mathcal{L}_{\text{Aleatoric}}$ (see Eq. 12), is not set statically, but adapted during the training process. More precisely, β_{KL} is set to zero for the first training epoch, allowing the optimisation process to focus on adapting the variational parameters with the exclusive objective of minimising the disparity error in the beginning of the training procedure. In the following five epochs, β_{KL} is incremented by 0.2 per epoch, gradually increasing the regularisation effect of the KL divergence. In all consecutive epochs, β_{KL} is constantly set to one. Finally, as stated in Equation 9, the training samples are weighted according to their disparity error, differentiating between values in three different ranges: β_{error} = 1.3 for a disparity error smaller than one pixel, β_{error} = 7.7 for a disparity error in the range of $[1, 5)$ pixels and β_{error} = 12.5 for a disparity error larger or equal than five pixels. The individual values of β_{error} are determined based on the error distribution of the training samples before starting to optimise for disparity and aleatoric uncertainty jointly.

4.3 Evaluation Strategy and Criteria

To set the results of the method proposed in this work in context and to allow a reasonable assessment, four different variants are examined and compared: *deterministic*, *deterministic + CVA-Net*, *probabilistic* and *probabilistic + CVA-Net*. *deterministic* is used as baseline and is equivalent to the original GC-Net proposed by Kendall et al. (2017). *deterministic + CVA-Net* complements the original deterministic GC-Net by CVA-Net as described in Section 3. *probabilistic* is equivalent to the method described in this work, but it is optimised for the estimation of disparity only, neglecting aleatoric uncertainty by replacing the term $\mathcal{L}_{\text{Aleatoric}}$ in Equation 12 with the L1 loss. Finally, *probabilistic + CVA-Net* is the complete method as proposed in this work. All four variants are trained following the strategy presented in Section 4.2. Lastly, according to (Mehltretter, 2020), the number of Monte Carlo samples K (cf. Eq. 2-4) that are drawn per test sample in the context of the two probabilistic variants is set to 50.

For the purpose of computing quantitative results, 100 random image pairs are used per dataset during testing (and all 15 image pairs in case of the Middlebury dataset) that have not been seen by the network, i.e., the training, validation and test sets are strictly separated. The disparity range considered in the experiments is adapted to each dataset based on the maximum ground truth disparity present in the respective dataset. The quality of the disparity estimates is measured using the Mean Absolute Error (MAE), the Root Mean Square Error (RMSE) and the Pixel Error Rate (PER). The PER is the percentage of pixels for which the difference between estimated and reference disparity exceeds a threshold τ, using one, three and five pixels as values for τ in the evaluation of this work. To assess the quality of the estimated uncertainty, the Pearson correlation coefficient $r_{\Delta d,\sigma}$ between the absolute disparity error Δd and the estimated uncertainty in form of the standard deviation σ is used.

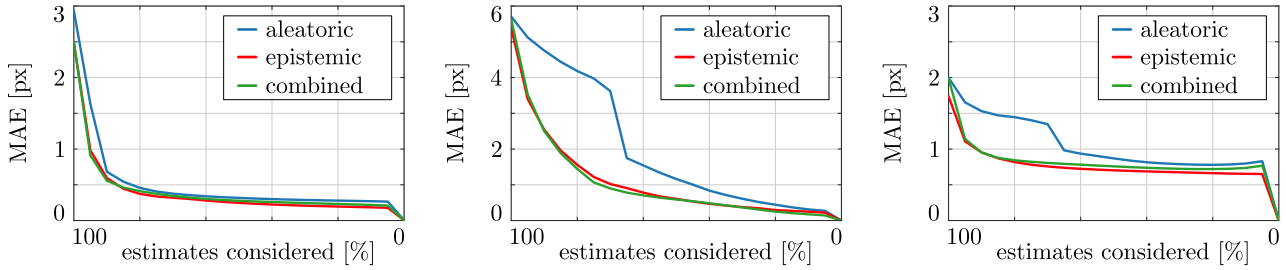
5. RESULTS

Analysing the correlation coefficients listed in Table 1, it can be seen that the variant that considers aleatoric and epistemic uncertainty jointly results in the highest correlation between the absolute disparity error and the estimated uncertainty for all three datasets evaluated. While the exclusive consideration of epistemic uncertainty leads to slightly worse results, only taking into account aleatoric uncertainty reduces the correlation significantly. It is also noticeable that the correlation decreases, with an increase of the domain gap between training and test data. While the correlations are highest on the InStereo2K dataset which was also used for fine-tuning the network parameters, they are worse for the Middlebury dataset, which also shows indoor scenes but with different characteristics and captured using a different set-up, and are worst for the KITTI dataset, which shows outdoor scenes and thus has the largest domain gap to the training data. In addition, the variant that only estimates aleatoric uncertainty seems to be especially sensitive regarding these differences in the data processed. This effect can be explained by the fact that such a domain gap is mainly reflected by the uncertainty embedded in the model, because the definition of domain gap implies that the statistical properties of the data used to train the parameters of a model differs from the properties of the data used to test this model. Consequently, as the uncertainty that is embedded in the model is neglected, the variant considering aleatoric uncertainty only is less suitable to estimate uncertainty that arises from a domain gap in the data.

These observations are also supported by the sparsification plots shown in Figure 2. In these plots, the mean absolute error is shown with respect to the percentage of disparity estimates considered, which is reduced discarding pixels having assigned the largest uncertainty estimates first. While all three variants lead to similar curves for the InStereo2K dataset, significant differences can be seen for the Middlebury and the KITTI dataset. For these two datasets, the exclusive consideration of aleatoric uncertainty is not sufficient to infer the disparity error from the uncertainty, leading to a clearly higher MAE for the same density compared to the two other variants. This behaviour is also illustrated by the qualitative examples shown in Figures 3 and 4. For the example from the InStereo2K dataset, the uncertainty estimates of all three variants allow to identify the majority of erroneous disparity estimates, most of them being part of an artefact located at the left side of the image which is caused by the complete absence of texture in this region. With respect to the example from the KITTI dataset, however, only the variants that consider epistemic uncertainty are capable of predicting uncertainty estimates that show a strong relation to the actual disparity error. In contrast, the uncertainty map obtained

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

Table 1. **Quantitative comparison of different variants of the method proposed.** The listed variants are analysed with respect to the disparity error metrics described in Section 4.3, the average of the estimated uncertainties, whereas the combined standard deviation is computed based on Equation 13, and the Pearson correlation coefficient of the absolute disparity error and the combined estimated standard deviation. A hyphen indicates that a certain type of uncertainty is not estimated using the respective model.



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

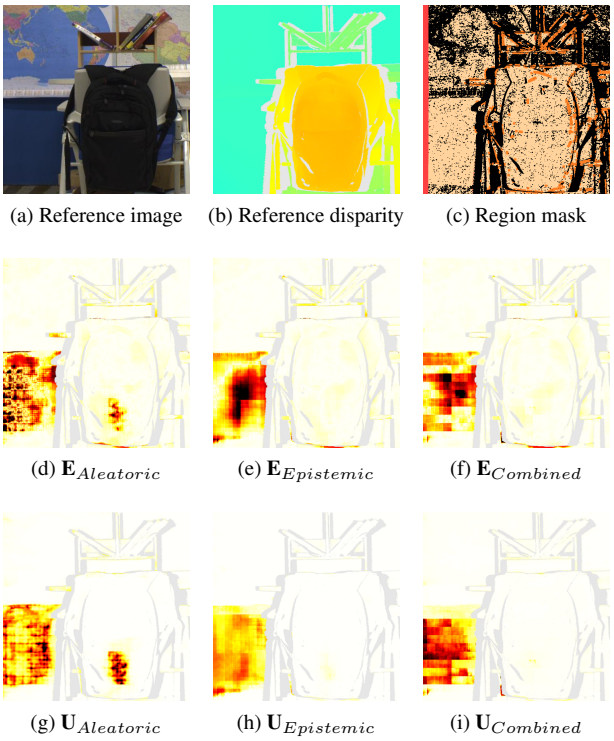
Figure 2. **Sparsification plots with respect to the different types of uncertainties estimated.** The figures show the mean absolute disparity error considering all test images of the respective dataset on the y-axis and the percentage of considered disparity estimates on the x-axis. The percentage is reduced discarding pixels that have assigned the highest uncertainties first. The depicted curves correspond to the aleatoric, epistemic and combined uncertainty predicted with the variants *deterministic + CVA-Net*, *probabilistic* and *probabilistic + CVA-Net*, respectively.

with the variant that considers aleatoric uncertainty only contains higher uncertainties for more distant points in the scene, but does not provide particularly large uncertainty estimates for pixels with a large disparity error.

Analysing the mean standard deviations listed in Table 1, it can be seen that the estimated aleatoric and epistemic uncertainty is always larger if only one type of uncertainty is considered in the estimation. This indicates that while aleatoric and epistemic uncertainty can be clearly separated in theory, to some extent both approaches are able to also account for uncertainty from sources assigned to the respective other type of uncertainty. However, the standard deviation of their combination is always larger than the individual uncertainties, implying that both types of uncertainty contribute to an accurate quantification and that a model that takes into account only aleatoric or only epistemic uncertainty is not capable to reflect the error distribution to be expected properly. As discussed before, this observation is also supported by the respective correlation coefficients.

The results corresponding to the disparity error reveal that introducing CVA-Net as additional network branch used to estimate aleatoric uncertainty does not only allow to assess the aleatoric uncertainty, but also influences the disparity estimation itself. While for the deterministic variant of GC-Net this influence can mainly be seen on the improved MAE and RMSE values for the InStereo2K dataset, the combination with CVA-Net has a positive effect on the pixel error rates for the probabilistic variant of GC-Net (cf. Tab. 1). While such an improvement can only be seen on the InStereo2K dataset that was used for fine-tuning the network parameters, the disparity error is slightly increased for the Middlebury and the KITTI dataset. This indicates that the combination of GC-Net and CVA-Net leads to an over-fitting to the characteristics of the training data, which has a negative impact on the transferability of a trained model to other datasets. Therefore, the results show that the joint estimation of disparity and aleatoric uncertainty has no negative impact on the disparity estimation capability if the training and the test data is similar, which demonstrates that this limitation stated in the literature can partly be overcome by the method proposed in this work. However, the observed over-fitting effect and thus the negative impact on the disparity estimates in the presence of a domain gap requires further investigations in future work.

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



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

(d) $\mathbf{E}_{Aleatoric}$ (e) $\mathbf{E}_{Epistemic}$ (f) $\mathbf{E}_{Combined}$

(g) $\mathbf{U}_{Aleatoric}$ (h) $\mathbf{U}_{Epistemic}$ (i) $\mathbf{U}_{Combined}$

Figure 3. **Qualitative comparison of different variants of the method proposed on an example of the InStereo2K dataset.** The figure shows the absolute disparity error maps $\mathbf{E}$ in comparison to the associated uncertainty maps $\mathbf{U}$ of the three variants that estimate aleatoric, epistemic and both kinds of uncertainty together, respectively. In both the error and the uncertainty maps, small values are shown in white, large ones in dark red / black. Note that the values of the three uncertainty maps are scaled to the same interval to allow for an easier comparison. The reference disparity map shows large disparities in orange to red and small ones in turquoise to dark blue. The region mask highlights regions that are especially challenging in the context of dense stereo matching, showing weakly textured areas in beige, occluded areas in red and pixels close to depth discontinuities in orange.

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

6. CONCLUSIONS

Addressing the task of uncertainty estimation in the context of dense stereo matching, a holistic approach is presented in this work that allows depth to be jointly estimated along with its associated uncertainty based on a stereo image pair. For this purpose, a BNN is proposed that is trained via Variational Inference, using a loss formulation that jointly optimises for the likelihood of a mixture distribution to estimate aleatoric uncertainty and for the similarity of the specified variational distribution and the exact posterior for the estimation of epistemic uncertainty. The experimental results underline the importance of estimating epistemic uncertainty: While the exclusive consideration of aleatoric uncertainty is sufficient to detect erroneous disparity estimates in the absence of a domain gap, it does not allow to capture the model uncertainty which typically dominates the uncertainty arising from the data given a strong difference in the characteristics of training and test data. Overall, the joint estimation of both, aleatoric and epistemic uncertainty, has demonstrated the best results and is thus the means of choice. The concepts for estimating aleatoric and epistemic uncertainty presented in this work, although only evaluated on the GC-Net architecture, can in principle be applied to any CNN architecture designed for dense stereo matching, requiring only the presence of some kind of cost volume. The practical applicability of both concepts in combination with more recent neural network architectures will be investigated in future work.

Besides the good results achieved, especially the analysis of the correlation between the disparity error and the estimated uncertainty reveals space for improvements. To keep the complexity of the assumed variational distribution low, a naive mean-field approximation with a Gaussian prior and a diagonal variance-covariance matrix is used as stochastic model for the proposed BNN in this work. Both are strong assumptions that potentially limit the quality of the estimated uncertainty. Consequently, further investigations on both, the definition of the prior and the consideration of correlations, for example, extending the mean-field approximation to a general formulation, are exciting directions for future research that promise further improvements.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159], the MOBILISE initiative of the Leibniz University Hannover and TU Braunschweig and by the NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Bao, W., Wang, W., Xu, Y., Guo, Y., Hong, S., Zhang, X., 2020. InStereo2K: A Large Real Dataset for Stereo Matching in Indoor Scenes. *Science China Information Sciences*, 63(212101).

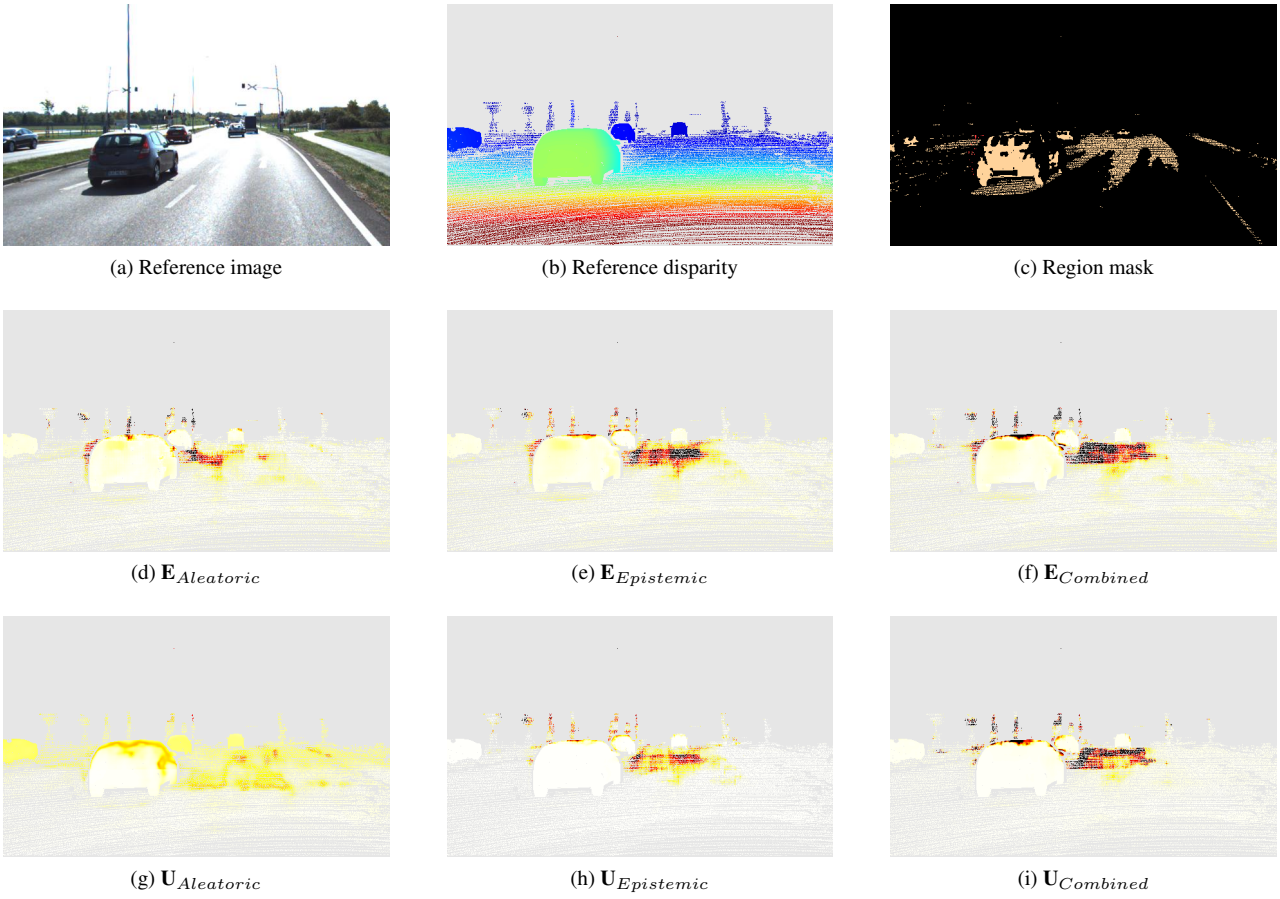
Brosse, N., Riquelme, C., Martin, A., Gelly, S., Moulines, É., 2020. On Last-Layer Algorithms for Classification: Decoupling Representation from Uncertainty Estimation. *arXiv preprint arXiv:2001.08049*.

Coenen, M., Rottensteiner, F., 2021. Pose Estimation and 3D Reconstruction of Vehicles from Stereo-Images using a Subcategory-aware Shape Prior. *ISPRS Journal of Photogrammetry and Remote Sensing*, 181, 27–47.

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

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

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



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

(d) $\mathbf{E}_{Aleatoric}$ (e) $\mathbf{E}_{Epistemic}$ (f) $\mathbf{E}_{Combined}$

(g) $\mathbf{U}_{Aleatoric}$ (h) $\mathbf{U}_{Epistemic}$ (i) $\mathbf{U}_{Combined}$

Figure 4. **Qualitative comparison of different variants of the method proposed on an example of the KITTI dataset.** The figure shows the absolute disparity error maps **E** in comparison to the associated uncertainty maps **U** of the three variants that estimate aleatoric, epistemic and both kinds of uncertainty together, respectively. For an explanation of the colour coding, refer to Figure 3.

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

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

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

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

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

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

Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in Neural Information Processing Systems*, 30, 5574–5584.

Kendall, A., Martirosyan, H., Dasgupta, S., Henry, P., Kennedy, R., Bachrach, A., Bry, A., 2017. End-to-End Learning of Geometry and Context for Deep Stereo Regression. *Proceedings of the IEEE International Conference on Computer Vision*, 66–75.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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