MIXED PROBABILITY MODELS FOR ALEATORIC UNCERTAINTY ESTIMATION IN THE CONTEXT OF DENSE STEREO MATCHING

Zeyun Zhong*, Max Mehltretter

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

Commission II, WG II/2

KEY WORDS: Uncertainty Quantification, 3D Reconstruction, Deep Learning, Mixture Model

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 two novel mixed probability models consisting of Laplacian and Uniform distributions 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.

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. It also serves as a foundation to support other photogrammetric computer vision problems, such as 3D reconstruction and object detection. 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 problems, such as occlusions, thin structures and large weakly textured areas, e.g. in the sky or caused by over-exposure. Consequently, the accuracy of the estimated depth information may be affected, making it 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. 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 two novel mixed probability models for aleatoric uncertainty estimation, using different combinations of a Laplacian and a Uniform distributions 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 mixture probabilistic aleatoric uncertainty model explicitly considering outlier measurements.

* Corresponding author

- An adaption of the CVA-Net architecture for the proposed uncertainty models. In addition, the network architecture is optimised to accelerate the training procedure.

2. RELATED WORK

As the ability to reliably detect failures of a stereo algorithm is fundamental, 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, 2016; 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, 2016) 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 in the form of the type and parameters of a particular distribution, uncertainty prediction has to be learned implicitly by assuming a specific uncertainty model. 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. 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 two novel uncertainty models based on two different assumptions, which are discussed in detail in Section 3.2. 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 carry out the task of uncertainty prediction, we test all variants using the same neural network architecture, which focuses only on the uncertainty estimation process. In detail, this architecture (briefly outlined in Section 3.1) utilises cost volumes as input, which are the result of the cost computation step of an arbitrary dense stereo matching approach carried out on an epipolar rectified image pair.

3.1 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) is utilised to evaluate our subsequently proposed probabilistic uncertainty models. This network consists of three major processing steps: First, a three-dimensional cost volume extract is merged to a single 1D feature vector, using 3D convolutional layers. To keep a good trade-off between the amount of information available to the network and the degree of smoothing within the resulting uncertainty map, we follow (Mehltretter and Heipke, 2021), setting the size of such extract to $13\times13\times192$. In the second step, the resulting 1D feature vector is further processed using 3D convolutional layers to derive high-level features characterising the cost volume extract. Based on the extracted features, two fully-connected layers, which are implemented in a convolutional manner, are then utilised to predict an uncertainty value in the last step.

While CVA-Net demonstrates convincing results, the usage of 3D instead of 2D convolutional layers leads to a significantly higher computational effort, decelerating the training process greatly. To mitigate this effect while maintaining comparable results, the kernel size of the 3D convolutional filters in the first part is changed from $3\times3\times3$ to $5\times5\times5$. In this way, 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 layer (Lin et al., 2014). 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. A detailed layer-by-layer definition of the modified architecture can be found in Table 1.

3.2 Probabilistic Uncertainty Models

Disparity estimation from stereo images is commonly learned in a supervised manner. However, reference data for the asso-

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 1. **Summary of the modified CVA-Net architecture.** Unless otherwise specified, each layer is followed by batch normalisation (BN) and a ReLU non-linearity. (Adapted from (Mehltretter and Heipke, 2021).)

ciated uncertainty, i.e. type and parameterisation of a particular probability distribution, is typically not available. Nevertheless, to 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 3.2.1, before two novel probabilistic models are proposed in Sections 3.2.2 and 3.2.3, 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.

3.2.1 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 estimated in advance and the predicted aleatoric uncertainty σ as the mean and 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,

$$-\log p(\hat{d}_i \,|\, d_i) \propto \frac{\sqrt{2}}{\sigma_i}|d_i - \hat{d}_i| + \log(\sigma_i), \tag{1}$$

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

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

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

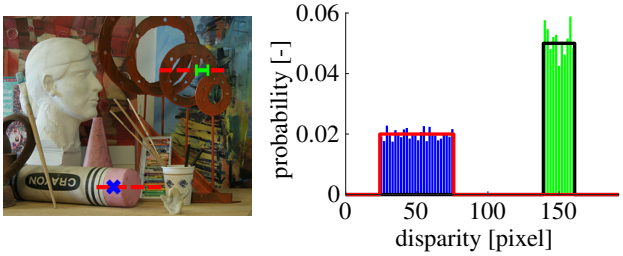


Figure 1. **Illustration of the underlying assumptions of the proposed geometry-aware model.** On the left, the right image of a stereo pair with a correct point match in a texture-less region and matching area in an occlued region are shown in blue and green, respectively. The considered search space 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.

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.

3.2.2 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, due to the geometry or appearance of the depicted scene.

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 (cf. Fig. 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. To better quantify the associated uncertainty for these pixels, such a disparity interval should be kept as tight as possible, containing only pixels which belong to the same object as the occluded one (cf. Fig. 1). In this work, 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. To overcome this limitation, a multimodal probability density function, for example, described as Gaussian mixture model, would need to be employed to describe the characteristics contained in these regions. However, this is beyond the scope of the present work and is subject of further investigations.

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 3.2.1, we consider the estimated disparity d as the mean of a Laplacian distribution 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:

$$L_{\mathcal{U}} = \begin{cases} 0.5x^2 & \text{if } |x| \leq \gamma \\ \gamma|x| - 0.5\gamma^2 & \text{otherwise,} \end{cases} \tag{3}$$

in which $x = |d_i - \hat{d}_i| - \sqrt{3}\,\sigma_i$. In this work, we set γ to 1, as commonly done in the literature (Mangasarian and Musicant, 2000; Girshick, 2015).

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:

$$L_{\mathcal{L}} = \frac{\sqrt{2}}{\exp(s_i)}|d_i - \hat{d}_i| + s_i. \tag{4}$$

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 the geometry-aware variant as:

$$L_{\text{Geometry}} = \frac{1}{N}\sum_{i=1}^{N} \hat{c}_i \cdot L_{\mathcal{L}} + (1-\hat{c}_i)\cdot L_{\mathcal{U}}, \tag{5}$$

where $\hat{c}$ denotes a binary parameter that is computed according to (Scharstein and Szeliski, 2002) and discussed in detail in Section 4.2. A pixel i gets assigned $\hat{c}_i = 1$ if it is located in a "good" region and 0 otherwise.

As the model described by Equation 5 does not learn to explicitly predict the type of region a pixel is assigned to, the ground truth region mask $\hat{c}$ would be required in the test phase to allow a proper interpretation of the predicted uncertainty. However, because the determination of occluded regions requires ground truth disparities, the model in its current form is not applicable to real-world applications. To overcome this limitation, we extend the loss function with a binary cross-entropy term $H(c, \hat{c})$, allowing the model to also learn the prediction of the region mask c. Thus, the loss function of the real-world applicable geometry-aware variant becomes:

$$L_{\text{Geometry*}} = \frac{1}{N}\sum_{i=1}^{N} \hat{c}_i \cdot L_{\mathcal{L}} + (1-\hat{c}_i)\cdot L_{\mathcal{U}} + H(c_i, \hat{c}_i), \tag{6}$$

$$H(c_i, \hat{c}_i) = -\hat{c}_i \cdot \log(c_i) - (1-\hat{c}_i)\cdot \log(1-c_i). \tag{7}$$

3.2.3 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. 3.2.2), 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 3.2.2, the loss function of our mixture model is defined as:

$$L_{\text{Mixture}} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i \cdot L_{\mathcal{L}} + (1-\alpha_i)\cdot L_{\mathcal{U}}, \tag{8}$$

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 (cf. Eq. 5 and Eq. 8). 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.

4. EXPERIMENTAL SETUP

In order to investigate the influence of the different uncertainty models (cf. Sec. 3.2), we train and evaluate these three models on three different datasets (discussed in detail in Section 4.1) 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 two metrics that are described in Section 4.4. 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 4.3.

4.1 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.

4.2 Binary Region Masks

For the training procedure of the geometry-aware uncertainty model discussed in Section 3.2.2, 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".

4.3 Training Procedure

Following the description in (Mehltretter and Heipke, 2021), we train the three 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 13×13×192 are extracted from normalised cost volumes corresponding to the left image of each pair. 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, cf. Tab. 1) 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 and 19 epochs for the Laplacian, the geometry-aware [1] and the mixture model, respectively.

4.4 Metrics

The first metric we use to evaluate the methodology presented in this paper, 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. The error rate is 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). The optimal AUC depends only on the overall error ϵ of a disparity map:

$$\begin{aligned} AUC_{opt} &= \int_{1-\epsilon}^{1} \frac{p-(1-\epsilon)}{p} dp \\ &= \epsilon + (1-\epsilon)\ln(1-\epsilon), \end{aligned} \tag{9}$$

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.

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 this estimates and the real disparity error. To overcome this limitation, we therefore use the correlation coefficient between the absolute disparity error and the estimated uncertainty as an additional metric to evaluate the uncertainty models discussed. The higher the correlation, the better the uncertainty can be quantified.

4.5 Evaluation Procedure

To ensure a clear separation of training and test data, we evaluate the three 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.

5. RESULTS

In this section, the results of two different sets of experiments are analysed and discussed. First, we investigate the applicability and compare the two variants of the proposed geometry-aware model introduced in Section 3.2.2. Subsequently, the ideal geometry-aware variant, i.e. the geometry-aware model with "perfect" region assignments, and the proposed mixture variant are compared against the state-of-the-art Laplacian variant in Section 5.2, in order to verify the validity of the proposed uncertainty models.

5.1 Analysis of the Geometry-aware Approach

To verify the applicability of the geometry-aware approach in real-world applications, the variant with region mask prediction is first compared with the ideal variant, which has "perfect" region assignments for all pixels, using the correlation coefficient.

[1] If not otherwise specified, in this paper the term geometry-aware model refers to the model predicting uncertainty only.

Correlation coefficient	Geometry		Geometry*	
	good	hard	good	hard
KITTI 2015 Menze and Geiger (2015)				
Census-BM	**0.84**	**0.80**	0.83	**0.80**
MC-CNN	**0.73**	**0.73**	0.69	0.72
Middlebury v3 Scharstein et al. (2014)				
Census-BM	**0.82**	**0.81**	0.79	0.78
MC-CNN	**0.71**	**0.78**	0.70	0.75

Table 2. **Comparison of the two variants of the proposed geometry-aware model based on the correlation coefficient.** The model marked with an asterisk jointly predicts an uncertainty map and a region mask, whereas the second variant predicts uncertainty values only. The 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.

As shown in Table 2, the variant with region mask prediction (marked with an asterisk) shows a slight deterioration of the uncertainty quantification, but the results are still comparable to those of the ideal variant. These differences are probably due to the increased complexity of the learning task caused by the additional consideration of the cross-entropy term in the loss function.

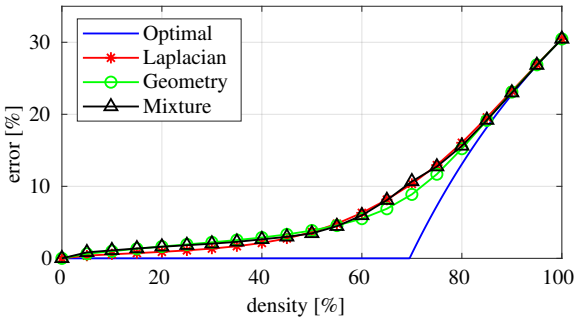
For a further investigation, the region masks predicted by the variant marked with an asterisk are evaluated using the overall accuracy (ACC), the true positive rate (TPR) and the true negative rate (TNR), with TPR and TNR describing the proportions of pixels that are correctly assigned to the "good" and "hard" regions, respectively. As shown in Table 3, the proportion of correctly assigned pixels is clearly higher in "good" regions than in "hard" ones in almost all configurations evaluated. Due to a noticeable domain gap for the configuration *MC-CNN + Middlebury v3*, the network is fine-tuned (see Tab. 2 for details) resulting in a TNR higher than the TPR for this configuration. This effect is probably caused by the higher percentage of "hard" samples in the Middlebury v3 dataset compared to the KITTI 2012 dataset, which is used to train all other variants and in which the majority of pixels are assigned to "good" regions. Consequently, the region mask prediction seems to be very sensitive to the imbalanced occurrence of classes in the training data. Addressing this class imbalance problem, a weighted cross-entropy term could be utilised to prevent the network from learning to preferably predict the more frequent class. Moreover, because the computation of the texture-less region masks is solely based on the stereo images which are also available during testing, these masks can be computed beforehand and do not need to be predicted by the network. Consequently, the region classification would be reduced to the identification of occluded pixels, which simplifies the learning task. Together with the prior knowledge with respect to the texture-less regions, the overall accuracy of the mask prediction is expected to be further improved. To conclude, the geometry-aware approach has shown its potential and applicability to real-world applications. However, further investigations are required that will be carried out in future work.

5.2 Uncertainty Model Evaluation

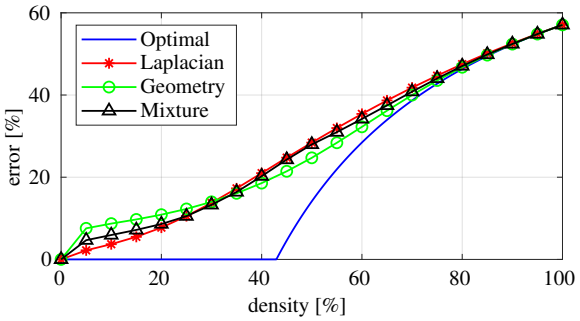
To verify the validity of the approaches proposed earlier in this work (see Sec. 3.2), the ideal geometry-aware and the mixture

[%]	KITTI 2015			Middlebury v3		
	ACC	TPR	TNR	ACC	TPR	TNR
Census-BM	80.69	89.94	59.26	74.99	89.92	62.85
MC-CNN	84.07	91.10	67.78	78.34	69.09	85.85

Table 3. **Accuracy (ACC), true positive rate (TPR) and true negative rate (TNR) of the region mask predictions of the real-world applicable geometry-aware model.** The TPR and TNR measure the proportions of pixels that are correctly assigned to the "good" and "hard" regions, respectively. For details on the evaluation procedure, please refer to Table 2.



(a) ROC curves for "good" regions



(b) ROC curves for "hard" regions

Figure 2. **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 4a.

variant are first analysed along with the state-of-the-art Laplacian variant with respect to the AUC metric (see Sec. 4.4). As shown in Table 4a, 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 (cf. Fig. 2). This observation confirms that these "hard" regions are especially challenging for dense stereo matching procedures, which is the fundamental motivation for our geometry-aware uncertainty model that explicitly accounts for these regions. On the other hand, the differences among the three uncertainty models with respect to the AUC values shown in Table 4a are not distinct. On all evaluated combinations of disparity methods and datasets the results of the three models are relatively similar. While the same can be observed in Figure 2, 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 4.4, the AUC metric neglects

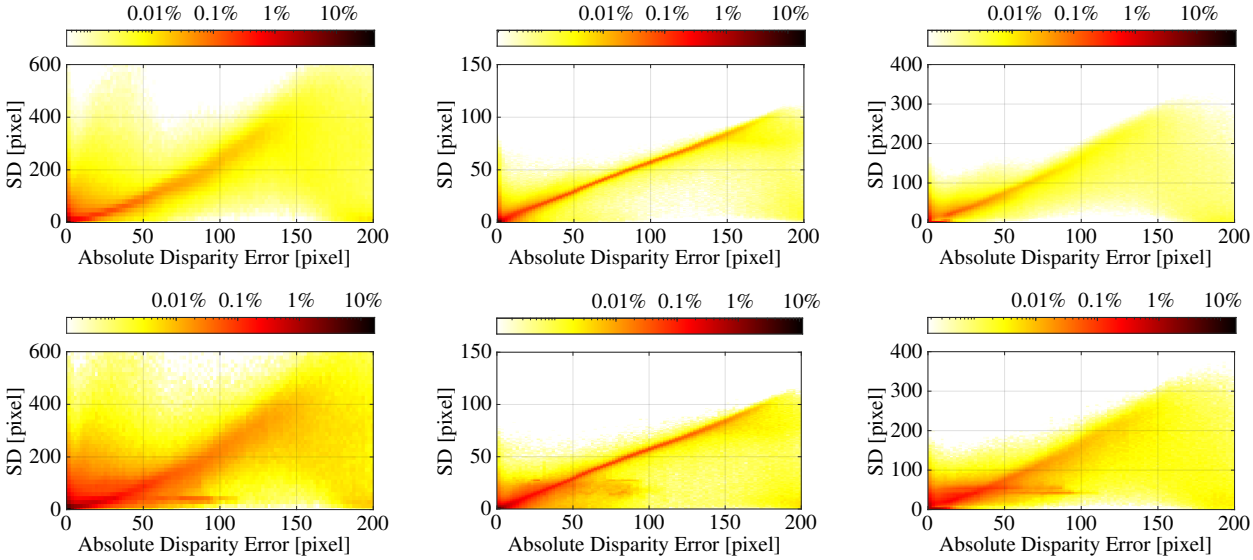
avg. AUC $= 10^{-2}\times$	Opt.		Laplacian		Geometry		Mixture	
	good	hard	good	hard	good	hard	good	hard
KITTI 2015 Menze and Geiger (2015)								
Census-BM	5.69	22.13	7.94	28.05	**7.92**	**27.83**	8.10	27.96
MC-CNN	1.06	7.19	**2.19**	10.35	2.40	11.04	2.27	**10.27**
Middlebury v3 Scharstein et al. (2014)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

(a) AUC comparison. The values represent the AUC$\times 10^{-2}$, whereas the smaller the values, the better, while Opt. is the best achievable value (cf. Sec. 4.4).

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

(b) Correlation coefficient comparison, measuring the correlation between the absolute disparity error and the estimated uncertainty. The bigger the values, the better.

Table 4. **Comparison of the three uncertainty models.** For details on the evaluation procedure, please refer to Table 2. While the three uncertainty models have only minor differences based on the AUC metric, the proposed geometry-aware model exceeds other models by a wide margin considering the correlation coefficient metric.



(a) Laplacian model (b) Geometry model (c) Mixture model

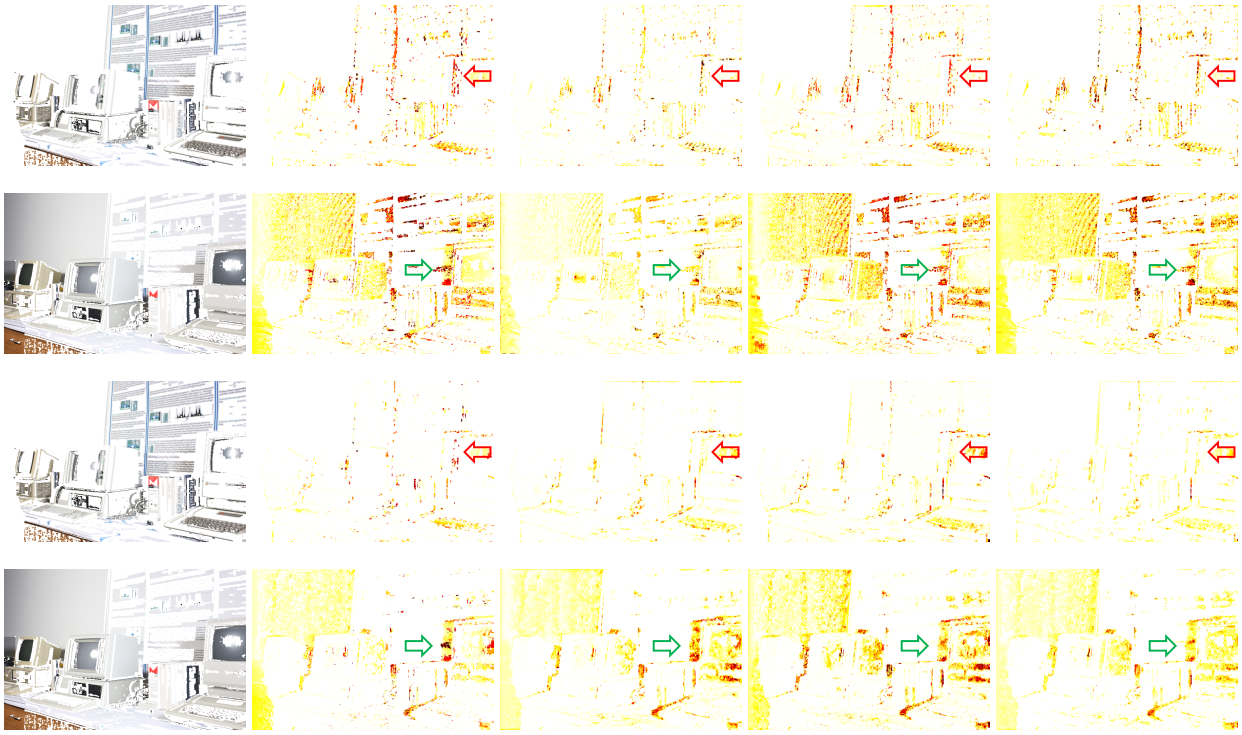
Figure 3. **Absolute error uncertainty relation.** The results are based on the first 100 images of the KITTI 2015 dataset using cost volumes computed with Census-based block matching. From top to bottom, the "good" and the "hard" regions are shown. The logarithmic colour scale encodes the percentage of pixels showing the respective error and estimated standard deviation (SD).

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. Analysing the correlation coefficients presented in Table 4b, it can be seen that the two 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 3: While the heatmaps of the Laplacian and the mixture model 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.

Figure 3 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 4: 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 two models proposed in this work (see Fig. 3). An example of this behaviour can be seen in the Census-based uncertainty maps corresponding to hard regions shown in Figure 5: 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.

According to the "good" regions in Figure 4, it can be seen that the uncertainties of pixels located at object boundaries (highlighted by a red arrow) are less accurate for all uncertainty models and for both disparity methods. This problem is caused by two potential reasons: First, the KITTI dataset, which was used for training, provides a sparse ground truth for the disparities

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-2-2021-17-2021 | © Author(s) 2021. CC BY 4.0 License.



(a) Reference images (b) Error maps (c) Laplacian model (d) Geometry model (e) Mixture model

Figure 4. **Qualitative uncertainty evaluation on the Middlebury v3 dataset.** From top to bottom, the "good" and the "hard" regions based on Census-BM and MC-CNN disparity methods are shown. 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 three uncertainty models.

and our network is trained in a supervised manner, only using patches centred on pixels with known ground truth. Consequently, edges in the disparity space caused by depth discontinuities are rarely seen by the network during training. Second, the assumed Laplacian or Laplace-Uniform mixture distribution is not suitable to describe the uncertainty related to depth discontinuities properly. In this case, it may be beneficial to utilise a multimodal distribution for the estimation of uncertainty at depth discontinuities. However, further investigations on that topic are necessary and will be carried out in future work.

6. CONCLUSION

In the present work, we propose two 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 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.

The results of the two proposed models demonstrate to be superior compared to the unimodal baseline, which is especially visible for occluded and texture-less areas of an image. However, the prediction of regions masks shows potential to be further improved, for example, by addressing the problem of imbalanced training data and the direct derivation of texture-less regions from the reference image. Moreover, the results have also shown that the two presented models as well as the baseline are less accurate for pixels close to depth discontinuities. To overcome this limitation, we plan to further investigate the possibilities of employing multimodal distributions, for example, in form of a Gaussian mixture model, to improve the estimation of aleatoric uncertainty in these areas.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) as a part of the Research Training Group i.c.sens [GRK2159] and the MOBILISE initiative of the Leibniz University Hannover and TU Braunschweig.

References

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

Fu, Z., Ardabilian, M., Stern, G., 2019. Stereo Matching Confidence Learning Based on Multi-modal Convolution Neural

(a) Reference images (b) Error maps (c) Laplacian model (d) Geometry model (e) Mixture model

Figure 5. **Qualitative uncertainty evaluation on the KITTI 2015 dataset.** From top to bottom, 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. 4. The uncertainty maps based on Census-BM generated by the two proposed models are smoother at the texture-less rear of the car, while the Laplacian model generates noisy uncertainties in this area.

Networks. *Representations, Analysis and Recognition of Shape and Motion from Imaging Data*, Springer, Cham, 69–81.

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.

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

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.

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

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

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

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

Mangasarian, O., Musicant, D., 2000. Robust Linear and Support Vector Regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 950-955.

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

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*, 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.

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

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

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

Scharstein, D., Hirschmueller, H., Kitajima, Y., Krathwohl, G., Nešić, N., Wang, X., Westling, P., 2014. High-resolution Stereo Datasets with Subpixel-accurate Ground Truth. *German Conference on Pattern Recognition*, Springer, 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), 7–42.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), 1929–1958.

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

Tosi, F., Poggi, M., Benincasa, A., Mattoccia, S., 2018. Beyond Local Reasoning for Stereo Confidence Estimation with Deep Learning. *Proceedings of the European Conference on Computer Vision*, 319–334.

Vogiatzis, G., Hernandez, C., 2011. Video-based, Real-Time Multi-View Stereo. *Image and Vision Computing*, 29(7).

Zabih, R., Woodfill, J., 1994. Non-Parametric Local Transforms for Computing Visual Correspondence. *Proceedings of the European Conference on Computer Vision*, Springer, 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), 2287–2318.