Uncertainty Estimation for Stereo Matching Based on Evidential Deep Learning

Chen Wang[a,**], Xiang Wang[a,**], Jiawei Zhang[a], Liang Zhang[a], Xiao Bai[a,*], Xin Ning[b,*], Jun Zhou[c], Edwin Hancock[d,a]

[a]*School of Computer Science and Engineering, State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University, China*
[b]*Institute of Semiconductors Chinese Adacamy of Sciences Laboratory of Artificial Neural Networks and High-speed Circuits, Beijing, China*
[c]*School of Information and Communication Technology, Griffith University, Nathan, Queensland 4111, Australia.*
[d]*Department of Computer Science, University of York, York, UK.*

Abstract

Although deep learning-based stereo matching approaches have achieved excellent performance in recent years, it is still a non-trivial task to estimate the uncertainty of the produced disparity map. In this paper, we propose a novel approach to estimate both aleatoric and epistemic uncertainties for stereo matching in an end-to-end way. We introduce an evidential distribution, named Normal Inverse-Gamma (NIG) distribution, whose parameters can be used to calculate the uncertainty. Instead of directly regressed from aggregated features, the uncertainty parameters are predicted for each potential disparity and then averaged via the guidance of matching probability distribution. Furthermore, considering the sparsity of ground truth in real scene datasets, we design two additional losses. The first one tries to enlarge uncertainty on incorrect predictions, so uncertainty becomes more sensitive to erroneous regions. The second one enforces the smooth-

*Corresponding author
**Equal contribution as first author

ness of the uncertainty in the regions with smooth disparity. Most stereo matching models, such as PSM-Net, GA-Net, and AA-Net, can be easily integrated with our approach. Experiments on multiple benchmark datasets show that our method improves stereo matching results. We prove that both aleatoric and epistemic uncertainties are well-calibrated with incorrect predictions. Particularly, our method can capture increased epistemic uncertainty on out-of-distribution data, making it effective to prevent a system from potential fatal consequences. Code is available at https://github.com/Dawnstar8411/StereoMatching-Uncertainty
Keywords: Stereo Matching, Uncertainty Estimation, Evidential Deep Learning

1. Introduction

Obtaining dense depth map is a crucial task in 3D reconstruction [1], visual SLAM [2], and autonomous driving [3]. Active 3D sensors, such as structured light, ToF cameras, and LiDAR, suffer from expensive imaging hardware, limited sensing range, or very sparse depth output. Stereo-based depth estimation is an alternative solution, which obtains dense disparity maps through stereo matching and then uses camera's imaging model to restore the depth of the scene. With the rapid development of deep learning technologies, many stereo matching models, such as PSM-Net [4], GA-Net [5] and AA-Net [6], have been proposed and achieved promising results. Despite the high performance, it is crucial to determine whether the model's output can be trusted, especially for some safety-critical applications. For example, for obstacle avoidance, which is a key feature in autonomous driving, it is not only necessary to obtain accurate depth information but also essential to know how reliable the predictions are. The stereo depth estimation model has a high possibility to fail when there are camera blurring, over-

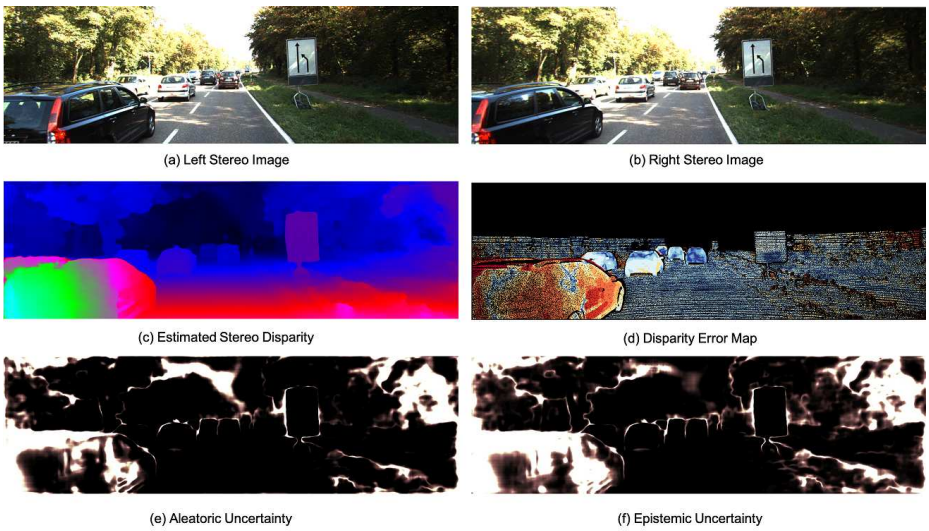

Figure 1: (a) Left stereo image. (b) Right stereo image. (c) Estimated disparity. (d) Disparity error map. Red means higher error. (e) Estimated aleatoric uncertainty that accounts for the regions which are hard to match, such as sky, object boundaries. (f) Estimated epistemic uncertainty that quantifies the uncertainty in the model. They are both well calibrated with erroneous regions.

exposure or an unfamiliar environment. Assigning high uncertainty to potentially wrong predictions can give warnings and prevent the autonomous driving system from making fatal decisions.

The uncertainties can be categorized as aleatoric or epistemic [7] depending on whether the error arises from imprecise data or poor knowledge. In stereo matching, aleatoric uncertainty is related to the input data, indicating regions that may be hard to match. Epistemic uncertainty captures the uncertainty in the model, which is suitable to identify the out-of-distribution data when the variations of the test domain are too complex to be covered in the training domain. Fig. 1 shows the visualized results of stereo matching and uncertainty estimation. Both aleatoric and epistemic uncertainties have high values on the erroneous regions.

In stereo matching, most approaches [8, 9] only model the aleatoric uncertainty by obtaining the confidence measure through handcrafted rules such as left-right consistency checking, but epistemic uncertainty is often ignored.

From the data analysis perspective, some approaches [10, 11, 12], such as Bayesian neural networks [10], consider both aleatoric and epistemic uncertainties. However, Bayesian neural networks place priors on network weights, which leads to a high computational sampling cost to estimate epistemic uncertainty during inference. Different from the Bayesian neural network, evidential approaches [13, 14, 15] consider learning as an evidence acquisition process. Priors are directly placed over the likelihood function to form a higher-order, evidential distribution. Training samples add evidence to fit this distribution. By learning to estimate the parameters of the evidential distribution, a grounded representation of both aleatoric and epistemic uncertainties can be obtained without the need for sampling. However, this strategy has not been applied to uncertainty estimation for stereo matching.

Two characteristics in stereo matching make it inappropriate to directly use evidential learning for uncertainty estimation. Firstly, stereo matching is not a strict classification [15] or regression [13] problem. In stereo matching networks [4, 5, 6], features of stereo image pairs are aggregated by 2D correlation or 3D cost volumes. In addition to the spatial features, the aggregated features also contain matching relationships. The process first calculates a D dimensional classification probability vector and then use it to obtain output disparity via soft-targmin. In the same way, it is not appropriate to directly regress uncertainty parameters from the aggregated features. In other words, the uncertainty should reflect the difficulty of matching, not just the final disparity result. The second

characteristic is that, in practice, it is extremely expensive and hard to capture dense disparity annotations, especially in autonomous driving. The uncertainty estimation model obtains sub-optimal results, lacking uncertainty constraints on regions without ground truth disparity.

To address the above-mentioned problems, we propose a novel approach that estimates uncertainty for stereo matching based on end-to-end evidential deep learning. Our method predicts uncertainties for each potential disparities at first and then takes stereo matching probabilities as guidance to softly average them to obtain the final uncertainty. Thus the matching behavior on each potential disparity contributes to the final uncertainty estimation. Furthermore, we propose two loss functions to constrain the uncertainty parameters despite the fact that there are no ground truth disparity annotations. The first loss minimizes evidence on incorrect predictions and inflates the uncertainty. The second loss constrains the smoothness of the uncertainties in regions with smooth disparities. By combining these two losses, pixels without ground truth disparities can be used in the training process to improve the performance of uncertainty estimation.

We utilize the Normal Inverse-Gamma (NIG) distribution as the evidential distribution. This models a higher-order probability distribution over the individual likelihood parameters. Given a stereo image pair, disparity, aleatoric and epistemic uncertainties can all be obtained through estimating the parameters of the NIG distribution. The backbone model can be replaced by most of the existing stereo matching networks with aggregated feature volume. We propose four branches to estimate the parameters respectively. The main contributions of this paper can be summarized as follows:

1) We propose a novel uncertainty estimation approach for stereo matching

based on evidential deep learning. Both aleatoric and epistemic uncertainties can be estimated in an end-to-end way. By utilizing matching probabilities as guidance to estimate uncertainty, the uncertainty well reflects the difficulty of matching.

2) We propose two loss functions to constrain the uncertainty parameters. By using prediction errors and a disparity agreement prior, pixels without ground truth disparity can be utilized during training to improve the performance of the uncertainty estimation.

3) We undertake comprehensive experiments to show that the proposed method improves stereo matching performance and obtains well-calibrated uncertainty. The proposed method not only assigns high uncertainty to erroneous estimation, but also captures increased epistemic uncertainty when there is out-of-distribution data.

2. Related Work

In this section, we give a brief overview of related work on deep learning based stereo matching and uncertainty estimation methods.

2.1. Deep Learning for Stereo Matching

Conventional stereo matching methods mainly consist of four stages: a) matching cost calculation, b) cost aggregation, c) disparity calculation/optimization, and d) disparity refinement [16]. Most recent stereo matching methods have leveraged deep learning approaches which considerably boost the accuracy of stereo matching thanks to their impressive feature representation capability. Early deep learning methods only utilized networks at some stages [17, 18, 19]. More recently, end-to-end neural networks for stereo matching have prevailed at processing and

are conceptually appealing with the availability of a relatively large amount of training data. These methods can be classified into 2D and 3D convolution based architectures depending on how the cost volumes are constructed. 2D architectures [20] use correlation to measure the similarity between features in the left image and their matching candidates in the right image (offset by disparity values). 3D architectures [4], on the other hand, directly concatenate the feature maps of the left image and their matching candidates and then let the network learn proper cost volumes.

PSM-Net [4] is a CNN based stereo matching method, containing a spatial pyramid pooling module and a 3D CNN. The spatial pyramid pooling module exploits global context information to form a cost volume and the 3D CNN learns to regularize cost volume. GA-Net [5] introduces a semi-global aggregation layer and a locally guided aggregation layer to capture both local and global cost dependencies. AA-Net [6] replaces the commonly used 3D convolutions with a sparse point-based intra-scale cost aggregation and a cross-scale cost aggregation module, which leads to fast inference speed. In this work, we adopt and compare these three networks as our backbone models.

2.2. Uncertainty Estimation Methods

Recent work mostly focuses on a) improving the accuracy [5, 21], b) designing faster and more efficient architectures [6], c) improving generalization and robustness via self-supervised learning [22] and d) domain adaptation [23]. In addition to these considerations, it is crucial to estimate the uncertainty of predicted results.

Uncertainty is a vital factor for safety-critical systems, as it gives a confidence measure associated with the estimation procedure adopted. An important issue for

stereo matching is to determine the confidence of the disparity map. Confidence values reveal the ranking of the reliability of the estimated disparity values among pixels and indicate possibly occluded pixels [8]. Thus, confidence measures can be utilized to indicate the pixels whose disparity values should be refined [18] or to aggregate predictions from different methods [24].

Recently, confidence measure estimation has benefited from deep learning and shown increased reliability of disparity estimation. Joint learning of disparity and confidence maps also improves the disparity map, since confidence values enable the detection and filtering of outliers. Reflective confidence networks [25] and unified confidence networks [26] jointly learn both confidence values and cost optimization to improve the final disparity estimation. Most recently, the disparity network and confidence network are trained jointly in an adversarial learning framework to make the confidence estimation method explicitly refine the disparity results in an end-to-end manner [27].

Althought confidence estimation methods can measure the reliability of stereo matching, they can not identify the sources of error. It is critical for matching methods, including deep learning models, to understand the regions or situations in which a model is uncertain about the estimated depth and the inherent reasons.

Kendall and Gal [10] first analyzed the uncertainty in computer vision applications and proposed a Bayesian deep learning framework for uncertainty quantification. Their framework considers two main sources of uncertainty. Firstly, aleatoric uncertainty accounts for the inherent noise in the observations. It is modeled as the variance of the Gaussian likelihood model learned via maximum likelihood training. Secondly, epistemic uncertainty quantifies the uncertainty in model parameters, which is obtained via Monte Carlo dropout (MC Dropout) dur-

ing inference as a variational Bayesian approximation [11]. An alternative to estimating epistemic uncertainty is to create an ensemble of multiple networks with random changes in the training setup, thus approximating the posterior distribution by the ensemble of several sampled distributions [28]. These methods have shown their efficiency and scalability in many computer vision tasks, such as classification (e.g. semantic segmentation) and regression (e.g. monocular depth estimation) [10]. However, multiple forward passes during inference are requested for obtaining epistemic uncertainty, resulting in a large consumption of both resources and time.

Several works have attempted sampling-free uncertainty estimation. Variance propagation methods [12] estimate epistemic uncertainty by injecting a noise layer into the network and approximating the variances at the output layer. Carvalho et al. calculated predictive uncertainty via functional variational inference and Gaussian processes [29]. Another avenue of investigaton is to place an explicit distribution over the distribution of the output values, specifically, over the hyperparameters of the output distribution. Two representative examples of such models are prior networks [14] and evidential methods [15], which are structurally similar but trained in a different manner. For classification, a Dirichlet distribution is placed over the softmax outputs and a network is trained to predict the hyperparameters of the Dirichlet distribution. These methods are effective in uncertainty estimation and out-of-distribution detection with a significant reduction of resource cost compared to both MC Dropout and ensemble approaches. Only very recently have these methods been applied to regression tasks such as monocular depth estimation. Deep evidential regression extends the evidential approaches by placing the evidential prior (a Normal Inverse-Gamma distribution) over the

Gaussian likelihood function [13]. This is closely related to our motivation of efficient and scalable uncertainty learning for stereo matching networks. Whereas, directly applying this work to stereo networks remains unexplored and deserves analyses in detail.

Uncertainty estimation has been widely applied to a variety of computer vision tasks. But most works only either model the aleatoric uncertainty or estimate the epistemic uncertainty using resource-demanding MC Dropout and ensembles. In the stereo matching task, Hu et al. [30] used predicted aleatoric uncertainty to refine disparity maps for high-resolution images. In this paper, we predict both aleatoric and epistemic uncertainties in a single forward pass for stereo matching. Evidential deep learning relies on ground truth disparity annotations and cannot be directly transformed to support unsupervised learning. We therefore introduce our work on uncertainty estimation for supervised stereo matching models.

3. Method

In this section, we describe our proposed uncertainty estimation approach for stereo matching. Fig. 2 illustrates the network architecture of our method. Given a stereo image pair, features are extracted and aggregated through CNN modules. Then the matching cost volume and uncertainty volume are predicted via four branches. Disparity, aleatoric and epistemic uncertainties can be obtained through the four estimated evidential distribution parameters γ, υ, α and β, which are calculated under the guidance of matching probability distribution. The network is trained by minimizing a hybrid loss function composed of evidential learning loss and two regularization terms. A detailed description of this method is given in the following subsections. Firstly, we briefly introduce the evidential deep

learning approach. Then we describe the details of our proposed loss functions. Finally, we describe the network architecture and how to obtain the parameters of the evidential distribution.

3.1. Evidential Deep Learning

Given two stereo images I_l and I_r, the stereo matching model learns to predict the disparity value d_i for each pixel in one of the stereoviews. The model can be optimized through the following absolute error loss:

$$L = \frac{1}{N}\sum_{i=1}^{N}|d_i - d_i^*|, \tag{1}$$

where N is the number of pixels and d_i^* is the ground truth disparity of the ith pixel. However, it does not explicitly model the underlying uncertainty in the data.

Uncertainty estimation can be performed on a maximum likelihood setting. We assume the estimated disparity d_i is drawn from a distribution, such as a Gaussian, with mean and variance parameters $\theta = (\mu, \sigma^2)$. By drawing observations y from the training data, model parameters can be learned to infer the parameter θ that maximize the likelihood $p(y|\theta)$. While the variance parameter σ^2 represents the aleatoric uncertainty, this method ignores the epistemic uncertainty.

Let us now consider the observations drawn from the Gaussian distribution but with unknown μ and σ^2. The unknown mean and variance, regarded as random variables, could follow the Gaussian distribution and the Inverse-Gamma distribution as their respective prior distributions:

$$\begin{gathered}(y_1, \ldots, y_N) \sim \mathcal{N}(\mu, \sigma^2) \\ \mu \sim \mathcal{N}(\gamma, \sigma^2 v^{-1}) \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta)\end{gathered} \tag{2}$$

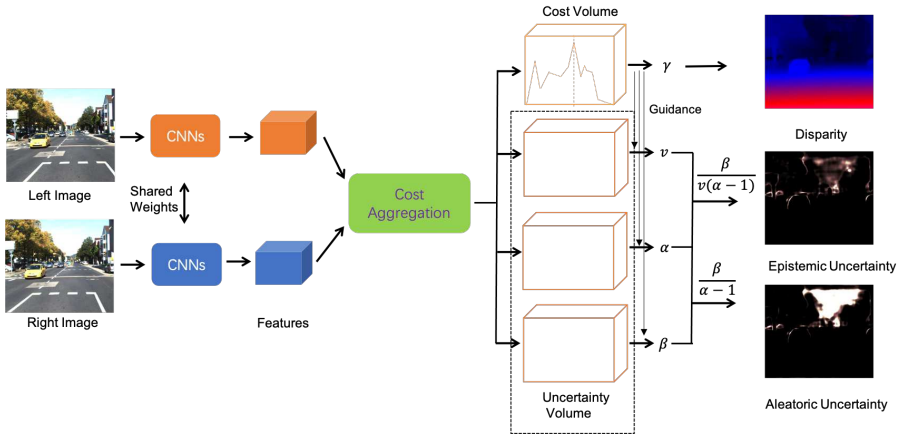

Figure 2: Illustration of the proposed uncertainty estimation approach for stereo matching. Features are extracted from input stereo image pair and cost aggregation is done by 2D or 3D CNN modules. Various cost aggregation methods can be used to aggregate features. Four parameters of the evidential distribution are estimated separately through the guidance of matching probability. The disparity, aleatoric uncertainty, and epistemic uncertainty can then be obtained.

where $\Gamma(\cdot)$ is the Gamma function, $\gamma \in \mathbb{R}, v > 0, \alpha > 1, \beta > 0$.

We want to estimate the posterior distribution

$$q(\mu, \sigma^2) = p(\mu, \sigma^2 | y_1, \ldots, y_N) \tag{3}$$

given observations from the available samples of training data. For computational tractability, we assume the independence of the mean and variance, which gives an factorized posterior

$$q(\mu, \sigma^2) = q(\mu) q(\sigma^2) \tag{4}$$

Thus, we can take the Normal Inverse-Gamma (NIG) distribution, i.e. the conjugate prior of the Gaussian distribution, as the approximated posterior. In deep evidential regression, Amini et al. [13] related this conjugate prior to evidential deep learning and defined the total evidence, $\Phi = 2v + \alpha$, to support the parame-

ter estimation.

Following the deep evidential regression framework, we train a network to infer the hyper-parameters $m = (\gamma, v, \alpha, \beta)$ of the NIG evidential distribution. Through the network, we can calculate both aleatoric and epistemic uncertainty along with disparity prediction in a single forward pass. Since the disparity value follows a Gaussian distribution, it can be expressed by the expectation of the mean of the Gaussian distribution. The aleatoric uncertainty indicates the degree to which the disparity value deviates from the ground truth. So it can be calculated by the expectation of the variance of the Gaussian distribution. The epistemic uncertainty represents the degree of dispersion of the disparity values. It can be calculated through the variance of the mean of the Gaussian distribution. The disparity, aleatoric and epistemic uncertainties are calculated as:

$$\mathbb{E}[\mu] = \gamma, \mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}, Var[\mu] = \frac{\beta}{v(\alpha - 1)} \tag{5}$$

To train the network that outputs the desired hyper-parameters m, we derive the loss term from the model evidence. The marginal likelihood thus maximizes model evidence in support of observations from the training data. The model evidence for the NIG prior and Gaussian likelihood follows a Student's t-distribution [13]. Based on Type II Maximum Likelihood, the loss function can be defined as the negative logarithm of the model evidence distribution:

$$\begin{aligned} \mathcal{L}_i^{data}(w) = &(\alpha + \frac{1}{2}) \log((y_i - \gamma)^2 v + \Omega) \\ &+ \frac{1}{2} \log(\frac{\pi}{v}) - \alpha \log(\Omega) + \log(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}) \end{aligned} \tag{6}$$

where $\Omega = 2\beta(1 + v)$. Amini [13] gives a detailed derivation for Eq. 6. As can be seen from the first term, when the error of predicted result becomes larger, the

evidence parameter α will become smaller. Thus incorrect results will correspond to high uncertainties. This loss function gives an objective to train the network to estimate evidential distribution parameter $m = (\gamma, v, \alpha, \beta)$ to with the support of traning samples by maximizing the model evidence.

3.2. Regularization Terms

In this section, we illustrate additional regularization terms to constrain the model via an incorrect prediction prior and a disparity agreement prior. The first term enforces lower estimated evidence parameters v and α in incorrect prediction regions, thus assigning high uncertainty on these regions. The second term constrains the smoothness of uncertainty values for regions with smooth disparity.

3.2.1. Regularization Based on Predictions

To ensure that the uncertainty and prediction errors are correctly calibrated, regularization terms should be introduced to minimize the weight of the evidence where the prediction is incorrect, while not influencing the evidence prediction where the prediction is close to the ground truth. A straight forward way to achieve the goal is to scale the total evidence Φ with the prediction errors as follows:

$$\mathcal{L}_i^{sup}(w) = |y_i - \mathbb{E}[\mu_i]| \cdot \Phi = |y_i - \gamma| \cdot (2v + \alpha) \tag{7}$$

This regularization term only works in a supervised setting since the error computation requires the ground truth disparity. However, dense ground truth is usually hard to obtain in real-world stereo matching. In the autonomous driving scenario, for example, the ground truth depth map is often acquired by LiDAR, which is both sparse and noisy. Evidence learning is regularized by the supervised term above where the ground truth is available, but it is not penalized where there

is no ground truth, which is a more frequent case. Therefore, appropriate loss terms are needed.

Here, we utilize the image reconstruction loss which is commonly used for self-supervised stereo matching methods. The ideal disparity map should ideally transform the right image of the stereo pair into the exact left image. Thus, a reconstructed left image I'_l can be obtained from the given corresponding right image I_r and the estimated disparity map d:

$$I'_l(x_i) = I_r(x_i - d_i) \tag{8}$$

The image reconstruction loss is used to maximize the photometric consistency between the original left image and the reconstructed left image, thus improving the quality of the predicted disparity map. A widely used image reconstruction loss is the weighted sum of the SSIM and the L1 loss:

$$\mathcal{L}_i^{pc}(w) = \lambda_{pc}\frac{1 - SSIM(I_l(x_i), I'_l(x_i))}{2} + (1 - \lambda_{pc})|I_l(x_i) - I'_l(x_i)| \tag{9}$$

However, this loss does not take uncertainty estimation into consideration. Following the original uncertainty estimation work [10], we define the regularized loss in the form of negative log-likelihood minimization for aleatoric uncertainty estimation. However we only take advantage of the property of learned loss attenuation, which forces model evidence negatively related to the image reconstruction loss. Considering the inverse of the total evidence, the loss is defined as:

$$\mathcal{L}_i^{epc}(w) = \frac{\mathcal{L}_i^{pc}(w)}{1/(\Phi - 1)} + \log(\frac{1}{\Phi - 1}) = (2v + \alpha - 1)\mathcal{L}_i^{pc}(w) - \log(2v + \alpha - 1) \tag{10}$$

Noting that $\Phi > 1$, here we use $\Phi - 1$ instead of Φ to ensure that $1/(\Phi - 1) \in \mathbb{R}^+$. Thus whatever the value of the image reconstruction loss, the model evidence gives reasonable values. The final form of the regularization term is as follows:

$$\mathcal{L}_i^{pred}(w) = \mathcal{L}_i^{sup}(w) + \mathcal{L}_i^{epc}(w) \tag{11}$$

3.2.2. Regularization Based on Disparity Agreement

We can penalize unreliable evidence for the whole image by extending the supervised regularization loss to the unsupervised learning setting. Nevertheless, since the image reconstruction loss provides a weaker training signal than supervised regularization, the output hyper-parameters of the NIG distribution are likely to contain a high level of noise, causing poor uncertainty estimation. Also, the above regularization only penalizes α and v, with the estimation of β not constrained. To reduce noise in the hyper-parameter estimation and make β constrained, we need to restrict the uncertainty of unannotated pixels by leveraging useful information from annotated pixels. We find that the hyper-parameter estimates for unannotated pixels can benefit most from the more reliable estimates from nearby annotated pixels. For the stereo matching task, this heuristic would be valid for most pixels except for those lying near the depth boundaries. Therefore, we propose a heuristic smoothness regularization on the output hyper-parameters β, α, v for pixels in smooth regions. Following the method proposed in [31], we rely on the disparity agreement between neighboring pixels to discard pixels close to depth boundaries, given by:

$$DA = \frac{\mathcal{H}_{N\times N}(d)}{N \times N} \tag{12}$$

where for each pixel $\mathcal{H}_{N\times N}$ denotes the number of neighboring pixels that have the same disparity value as the center pixel (considering sub-pixel precision within

1 pixel). Clearly, the higher the disparity agreement values are, the smoother the disparity patches behave. Thus, we define a mask M that contains pixels with high DA (heuristically, $M = \mathbb{I}(DA > 0.5)$, where $\mathbb{I}$ is the indicator function), representing whether pixels lie in smooth regions. Those pixels have similar hyper-parameter values to their neighbors. Therefore, we introduce a smoothness regularization loss that penalizes the hyper-parameter gradient variation on the smooth regions masked by M:

$$\mathcal{L}_i^{smooth}(w) = M_i(|\partial_x \alpha_i| + |\partial_y \alpha_i|) \tag{13}$$

Here ∂_x and ∂_y are the horizontal gradient and vertical gradient respectively. Similar smoothness regularization can be defined for β and v. With the help of the proposed smoothness regularization, the training of a hyper-parameter network can be more constrained, especially when the ground truth annotations are sparse and noisy. To summarise, the total loss function is the weighted sum of all the above loss terms and is given by:

$$\mathcal{L}_i(w) = \lambda_{sup}\mathcal{L}_i^{data}(w) + \lambda_{pred}\mathcal{L}_i^{pred}(w) + \lambda_{smooth}\mathcal{L}_i^{smooth}(w) \tag{14}$$

where $\lambda_{sup}, \lambda_{pred}, \lambda_{smooth}$ are the corresponding loss weighting factors.

3.3. Network Architectures

Fig. 2 shows the detailed architecture of our proposed approach. Most deep stereo matching networks, such as PSM-Net [4], GA-Net [5] and AA-Net [6], can be adopted as backbone stereo matching network and easily integrated with our proposed approach. In these stereo matching networks [4, 5, 6], given a stereo image pair, siamese networks with shared weights are used to extract features

and cost aggregation module is used to aggregate feature to form a feature volume. Our proposed method keeps the feature extraction and aggregation parts unchanged. Additionally, the proposed method has four output branches, one for each of the hyperparameters of the NIG evidential distribution.

The first branch estimates disparity values (the parameter γ). For each pixel, we firstly calculate the classification probability for D candidate disparity values $(1, 2, ..., 192)$ and then use softargmax operation to obtain the final disparity, which is defined as:

$$softargmax := \sum_{1}^{D} d * \delta(p_d) \tag{15}$$

where $d \in [1, D]$ is candidate disprity value, $\delta(\cdot)$ is the $softmax$ operation, and p_d is the normalized classification probability on disparity d. The parameters β, α and v are obtained via the remaining three branches, respectively. For each parameter, analogous to the first branch, we learn D hypotheses from the aggregated features, each corresponding to a candidate disparity. Note that the classification probability distribution after softmax denotes the probability that a certainty disparity candidate would be chosen, we use it as a hypotheses selector to select the most likely parameter. Since the argmax operation is not differentiable, we use the weighted sum of parameter hypotheses where the weight is given by the probability distribution.

4. Experimental Results

In this section, we evaluate our proposed approach. Firstly, we introduce implementation details of both training and evaluation procedures to ensure full reproducibility. Stereo matching and uncertainty estimation results are then analyzed. We also conduct two ablation studies.

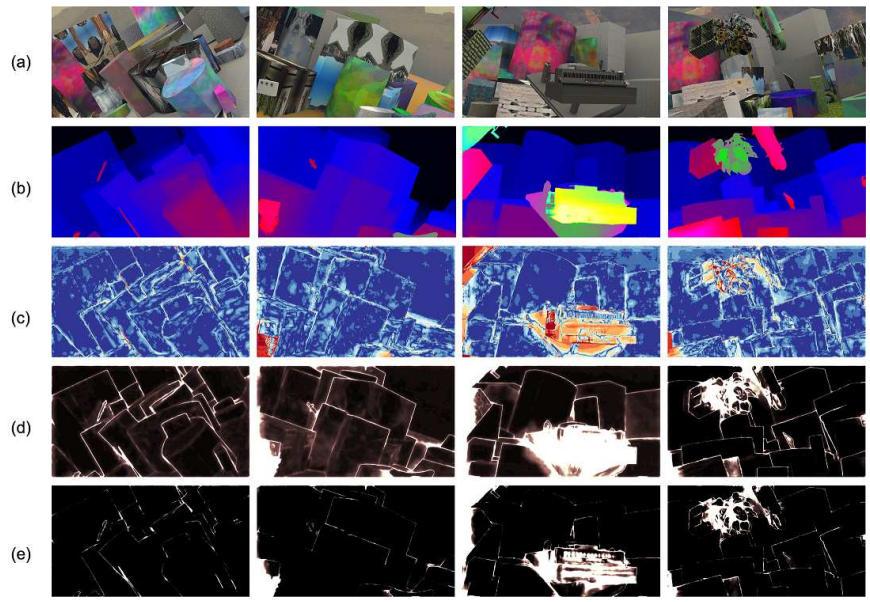

Figure 3: (a) Left stereo image. (b) Estimated disparity through our proposed approach. (c) Disparity error map. (d) Aleatoric uncertainty. (e) Epistemic uncertainty.

4.1. Implementation Details

Datasets. Three benchmark stereo matching datasets are employed for all experiments in this work. Sceneflow [32] is a large-scale synthetic dataset for the evaluation of stereo methods. It contains over 26k stereo image pairs which are split into two parts: 22k image pairs for training and 4k image pairs for testing. Sceneflow contains sub-pixel level dense ground truth disparity maps, with also a high diversity of different scenes. The maximal disparity of each image ranges from 20 to over 1000. During training, we filter out pixels whose ground truth disparity is greater than 192 and only calculate losses for the rest of the pixels. KITTI [33, 34] is a real scene stereo dataset captured by vehicle-mounted stereo cameras and annotated by LiDAR mounted behind the left camera. Different from the synthetic dataset, it only provides sparse disparity maps with about 30% pixels

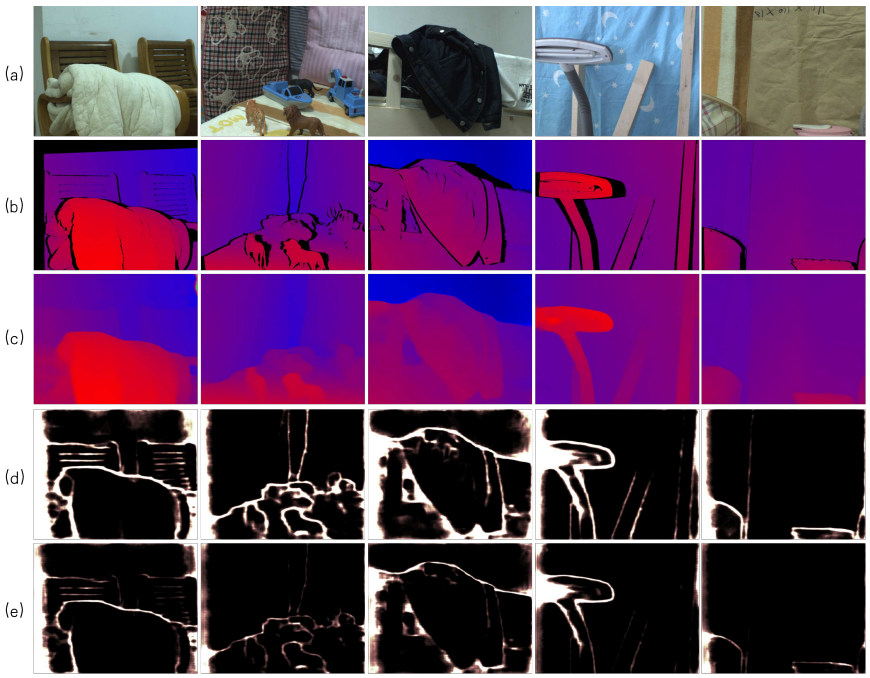

Figure 4: (a) Left stereo image. (b) Ground Truth Disparity. (c) Estimated disparity through our proposed approach. (d) Aleatoric uncertainty. (e) Epistemic uncertainty.

annotated for each image. Containing various real street scenes, this dataset raises significant challenges for stereo matching algorithms. KITTI contains two versions of datasets: KITTI2012 and KITTI2015. In this work, we mix all training images of KITTI2012 and 120 training images of KITTI2015 as the training set. We use the rest 80 training images of KITTI2015 for evaluation. InStereo2K [35] is a real dataset for stereo matching in indoor scenes. It contains 2000 pairs of stereo images for training and 50 pairs of stereo images for testing. Each pair of images corresponds to high accuracy disparity map.

Training Procedure. As our proposed approach is suitable for most stereo matching methods, we choose PSM-Net [4], GA-Net [5] and AA-Net [6] as the

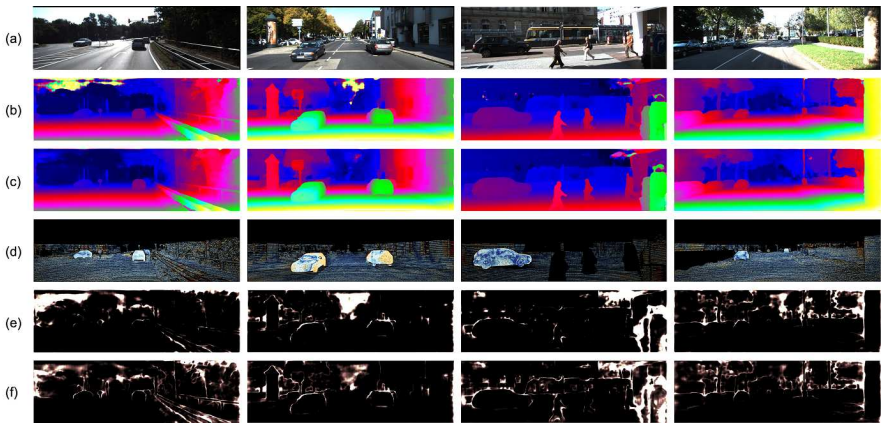

Figure 5: (a) Left stereo image. (b) Estimated disparity through the original PSM-Net. (c) Estimated disparity through our proposed method. (d) Disparity error map. (e) Aleatoric Uncertainty. (f) Epistemic Uncertainty.

base model. The source code is implemented using PyTorch framework. For all experiments, stereo image pairs are randomly cropped and fed into the network during training. The size of cropped image is consistent with the number in the original paper of the base model. Using a batch size of 1, the network is trained on 4 NVIDIA 2080TI GPUs. The weight parameters of the network are initialized using a uniform distribution and optimized using the Adam algorithm. We pre-trained the network on the Sceneflow dataset for 15 epochs. The initial learning rate is set as 0.001 when training with Sceneflow dataset and decreased by a factor of 10 after 10 epochs. The network is fine-tuned on KITTI dataset for 200 epochs. We set the initial learning rate as 0.001 and decrease it to 0.0001 after 100 epochs. By tuning the balance parameters in training loss, the final values are identified as $\lambda_{sup} = 1, \lambda_{pred} = 1, \lambda_{smooth} = 0.1$.

Evaluation Metrics. We use mean absolute error (MAE) as the metric to evaluate the average deviation of the estimated disparity from the ground truth.

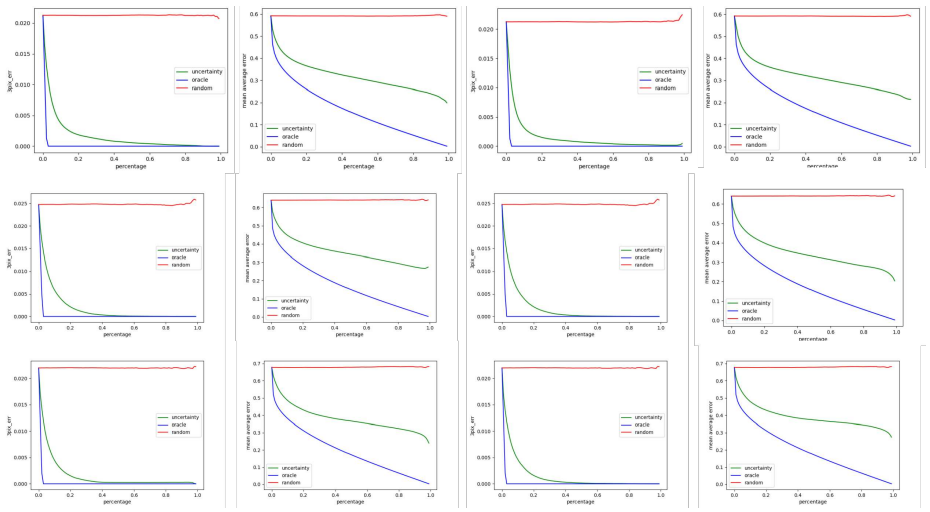

Figure 6: Sparsification plots. The red, blue and green lines means random plot, oracle plot and uncertainty plot. From tow to bottom are the sparsification plots of our proposed method, Bayesian deep learning method and bootstrapped method. From left to right are Aleatoric uncertainty sparsification plot under 3 pixel error metric, Aleatoric uncertainty sparsification plot under mean average error metric, Epistemic uncertainty sparsification plot under 3 pixel error metric and Epistemic uncertainty sparsification plot under mean average error metric.

Moreover, we also adopt the percentage of "bad" pixels with different thresholds 2, 3,5 as usually reported in the literature [22].

To evaluate the quality of estimated uncertainties, we adopt commonly used sparsification plots. All pixels in the disparity map are sorted in order of descending uncertainty. Then, the pixels with the highest uncertainty are removed gradually and the disparity metrics are calculated on the remaining pixels. If the estimated uncertainty properly represents the errors in the disparity map, the curve should be decreasing. The best sparsification plot is obtained by ranking pixels according to the true errors. We call this curve oracle sparsification. In contrast, a random uncertainty estimation leads to a flat curve because it offers no use-

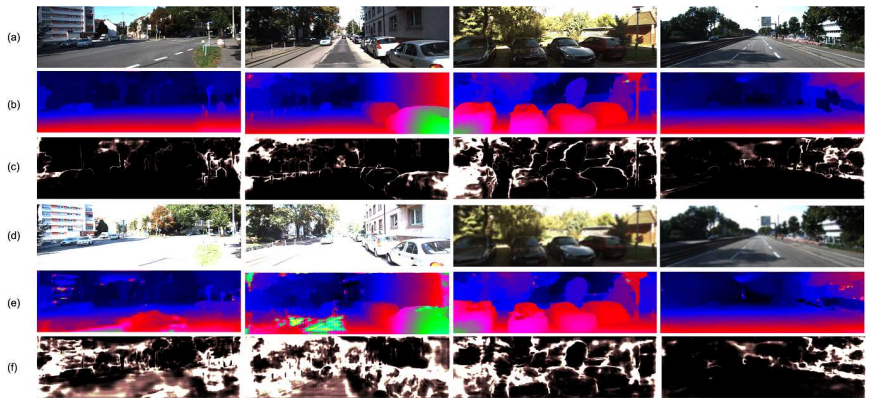

Figure 7: (a) Left stereo image. (b) Estimated disparity under normal settings. (c) Epistemic uncertainty under normal settings. (d) Left stereo image with overexposure or blurring. (e) Estimated disparity under settings(d). (f) Epistemic uncertainty under settings(d).

ful information about which pixels are bad. We adopt two quantitative metrics: Area Under the Sparsification Error (AUSE) and Area Under the Random Gain (AURG). AUSE means the difference between estimated and oracle sparsification, thus lower is better. AURG is obtained by subtracting estimated sparsification from a random one, so higher is better.

4.2. Disparity Estimation results

We evaluate the performance of our proposed methods on stereo matching. PSM-Net [4], GA-Net [5] and AA-Net [6] are used as the backbone model. We design Three kinds of training and testing splits. The first one trains and tests the models only on Sceneflow dataset. The second one pre-trains the models on Sceneflow dataset and fine-tune it on KITTI dataset.The models are tested on KITTI dataset. The third one pre-trains the models on Sceneflow dataset and fine-tune it on Instereo dataset. The models are tested on Instereo2K dataset. We compare the stereo matching performance of our proposed approach with the orig-

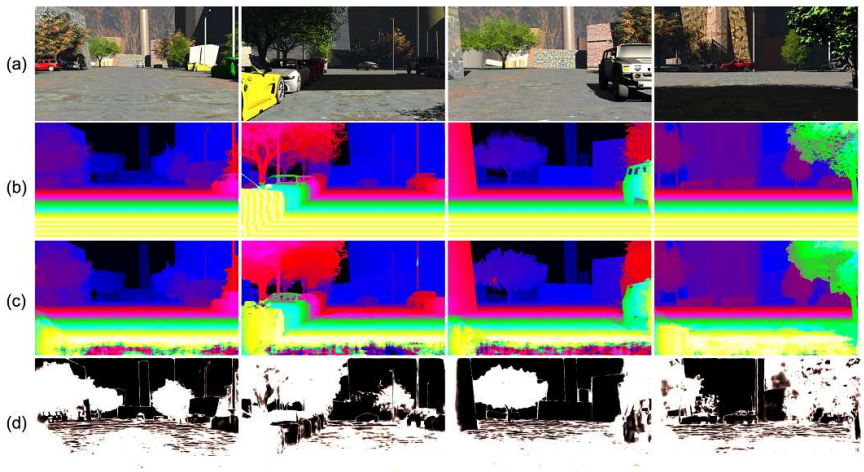

Figure 8: (a) Left stereo image. (b) Ground truth disparity. (c) Estimated disparity. (d) Epistemic Uncertainty. Even through the synthetic image has the similar street scene as KITTI dataset, the model still assigns high epistemic uncertainty.

inal stereo matching models. Table 1 shows that modeling aleatoric and epistemic uncertainties via our proposed method can improve the stereo matching performance. In stereo matching, some ill-posed regions, such as occlusion and reflection, are hard to match. Modeling uncertainties can reduce the negative effect of the ill-posed regions with the implied attenuation during training. In Eq. 6, the network tends to predict smaller alpha values which correspond to high uncertainties. A small alpha will reduce the negative effect of ill-posed regions. On KITTI dataset, the improvement is a little more pronounced. Fig. 3, Fig. 4 and Fig. 5 show the visualized results of stereo matching and uncertainty estimation. Fig. 5 (b) is the disparity estimated from the original PSM-Net. Fig. 5 (c) is the disparity estimated through our proposed approach. Some regions are hard to match, such as sky, object boundaries, and thin objects. Our proposed method performs better in these regions. We also observe that high uncertainties are assigned on dis-

Table 1: Quantitative results of disparity estimation."≥2 px", "≥3 px" and "≥5 px" show the percentage of pixels that have more than two, three and five pixels disparity error respectively.

Method	Training Datasets	Test Datasets	Error(%)			Mean Error
/		KITTI	≥2 px	≥ 3 px	≥ 5px	
PSM-Net	SceneFlow	SceneFlow	5.23	3.83	2.45	0.88
PSM-Net	SceneFlow+KITTI	KITTI	3.86	2.47	1.87	0.64
PSM-Net	SceneFlow+Instereo2K	Instereo2K	4.32	3.41	2.28	0.77
PSM-Net-un	SceneFlow	SceneFlow	5.11	3.43	2.31	0.82
PSM-Net-un	SceneFlow+KITTI	KITTI	3.78	2.17	1.75	0.59
PSM-Net-un	SceneFlow+Instereo2K	Instero2K	4.26	3.35	2.19	0.73
GA-Net	SceneFlow	SceneFlow	5.45	3.75	2.31	0.84
GA-Net	SceneFlow+KITTI	KITTI	3.91	2.54	1.86	0.65
GA-Net	SceneFlow+Instereo2K	Instereo2K	4.37	3.48	2.21	0.79
GA-Net-un	SceneFlow	SceneFlow	5.12	3.43	2.23	0.80
GA-Net-un	SceneFlow+KITTI	KITTI	3.84	2.34	1.72	0.62
GA-Net-un	SceneFlow+Instereo2K	Instero2K	4.31	3.39	2.15	0.74
AA-Net	SceneFlow	SceneFlow	5.67	3.80	2.39	0.87
AA-Net	SceneFlow+KITTI	KITTI	3.98	2.60	1.88	0.68
AA-Net	SceneFlow+Instereo2K	Instereo2K	4.26	3.29	2.21	0.73
AA-Net-un	SceneFlow	SceneFlow	5.56	3.67	2.28	0.85
AA-Net-un	SceneFlow+KITTI	KITTI	3.87	2.36	1.78	0.64
AA-Net-un	SceneFlow+Instereo2K	Instero2K	4.14	3.16	2.03	0.70

tant objects and on object boundaries which have a high probability to get wrong predictions. Modeling uncertainty in stereo matching can be used to learn loss attenuation and thus improve the accuracy.

The estimated uncertainty can be used to rectify the "uncertainty" region. We applied two kinds of "refinement mechanisms". First, we add a CNN-based refinement network that takes disparity map and uncertainty map as input. The output is considered as residual disparity and added to the original disparity. We use PSM-Net as backbone and evaluate the performance on SceneFlow dataset.

Table 2: Quantitative results of uncertainty estimation. AUSE means Area Under the sparsification error (lower is better). AURG means Area Under the Random Gain (higher is better). Mean Error and "≥3 px" are adopted disparity evaluation metrics.

Method	PSM-Net				GA-Net			
/	Mean Error		≥ 3 px		Mean Error		≥ 3 px	
/	AUSE	AURG	AUSE	AURG	AUSE	AURG	AUSE	AURG
Bayesian	0.182	0.267	0.00119	0.0173	0.178	0.276	0.00116	0.0174
Ensembel	0.179	0.271	0.00117	0.0176	0.176	0.278	0.00115	0.0172
Ours	0.174	0.278	0.00114	0.0183	0.173	0.281	0.00113	0.0176

Table 3: Quantitative results of architecture estimation.

Method	Disparity		Uncertainty			
/	Mean Error	≥ 3px	Mean Error		≥ 3px	
/	/	/	AUSE	AURG	AUSE	AURG
PSM-Net-re	0.68	2.57	0.1823	0.2713	0.001201	0.0178
PSM-Net-un	0.59	2.17	0.1739	0.2782	0.001137	0.0183
GA-Net-re	0.69	2.64	0.1789	0.2767	0.001189	0.0193
GA-Net-un	0.62	2.34	0.1728	0.2814	0.001130	0.0179
AA-Net-re	0.72	2.64	0.1802	0.2865	0.001278	0.0172
AA-Net-un	0.64	2.36	0.1765	0.2912	0.001204	0.0180

The mean error dropped from 0.88 to 0.85. Second, we consider the uncertainty as a mask and replace the disparity with high uncertainty using disparity of surrounding pixel with lower uncertainty. The mean error dropped from 0.88 to 0.86.

4.3. Uncertainty Estimation results

We compared with two state-of-the-art uncertainty estimation methods. Bootstrapped ensemble methods train an ensemble of N randomly initialized neural networks. During the inference time, they combined all model predictions to cal-

Table 4: Quantitative results of loss estimation.

Method	Disparity		Uncertainty			
/	Mean Error	$\geq$ 3px	Mean Error		$\geq$ 3px	
/	/	/	AUSE	AURG	AUSE	AURG
loss-1	0.63	2.40	0.1801	0.2734	0.001181	0.0180
loss-2	0.61	2.24	0.1768	0.2746	0.001154	0.0181
loss-3	0.02	2.25	0.1778	0.2743	0.001163	0.0181
loss-4	0.59	2.17	0.1739	0.2782	0.001137	0.0183

culate the epistemic uncertainty. Bayesian deep learning-based uncertainty quantification methods use dropout strategy to estimation uncertainty. Table 2 summarizes the results of the estimated uncertainties. PSM-Net [4] and GA-Net [5] with uncertainty estimation are trained on Sceneflow dataset at first and fine-tuned on KITTI. We evaluate the results on KITTI dataset. Table 2 shows AUSE and AURG measures of these methods. Our proposed method obtains the highest performance. Our proposed method trains the network in an end-to-end way and predict the disparity and uncertainty parameters in one forward pass during inference. Our method efficiently reduce the time and computation consumption. Fig. 6 visualizes the sparsification plots. By removing pixels with large uncertainty, the model performance improves. It shows that the aleatoric and epistemic uncertainties can be well-calibrated with wrong predictions.

We also evaluate the ability of our proposed uncertainty estimation method to capture increased epistemic uncertainty on out-of-distribution data. We choose three representative situations, including blurring, overexposure, and dataset variation. We use Gaussian kernel to blur the stereo images. Gamma operation is used to simulate different light conditions. As our model is trained on KITTI dataset,

we also test it on synthetic street scenes. Fig. 7 shows the disparity and epistemic uncertainty estimation results. Under strong light or image blurring, the quality of stereo matching decreases. The epistemic uncertainty values have increased correspondingly. In autonomous driving cases, it is significant to capture these special circumstances and give warning information. Fig. 8 shows the disparity and epistemic uncertainty estimation results on synthetic street scenes. Since they have the same street scenes with KITTI dataset, the performance of disparity estimation does not degrade too much. Our proposed method still assigns high epistemic uncertainties to show that they are out-of-distribution samples.

We compared with three confidence estimation methods [36, 26, 27]. We use a well-trained PSM-Net to obtain the raw cost volume and disparity map as the input of confidence estimation models, and then train the confidence estimation model. We evaluate the performance of confidence estimation using AUSE and AURG measures under 3 pixels error rates on Sceneflow database. We compare the aleatoric estimation results with these confidence prediction methods. The AUSE and AURG values of these three confidence estimation methods are 0.00117/0.0180, 0.00116/0.0181, 0.00115/0.0181. The AUSE and AURG values of our proposed methods are 0.00114/0.0183. These methods have comparable performance. However, our proposed method aims to predict different uncertainties and is an end-to-end way to estimate both disparity and uncertainty. It has some advantages over the confidence estimation methods.

4.4. Ablation Study

Architecture. We compare two kinds of network architectures to estimate the parameters of the evidential distribution. The first one uses convolution to directly regress v, α, and β from the output of the cost aggregation module. The matching

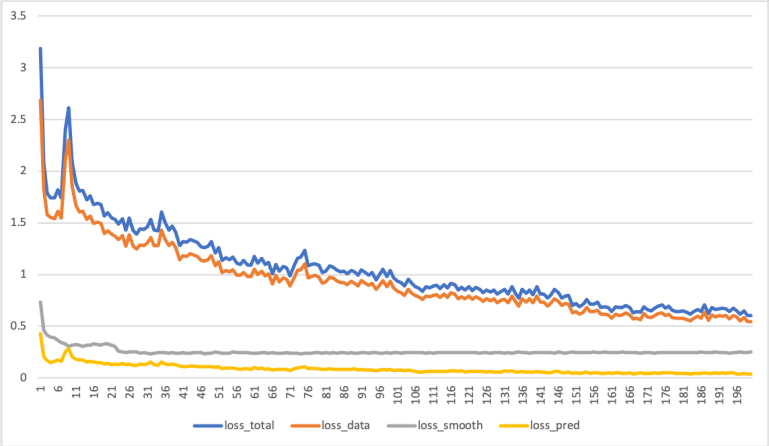

Figure 9: Learning curves of loss functions.

probability is not used. We use PSM-Net-re, GA-Net-re and AA-Net-re to represent them. The second one is our proposed approach. Uncertainty parameters are predicted for each potential disparity at first. The matching probability is used as the guidance to obtain output evidence through a weighted average operation. Table 3 shows that the first architecture performs worse on uncertainty estimation. Moreover, it reduces the accuracy of stereo matching, which is not what we wanted. In the absence of matching probabilities as a guide, the same convolution kernel for all pixels cannot distinguish evidence on different disparities. By useing matching probability as the guidance, the uncertainty effectively reflects the difficulties of stereo matching.

Loss. We propose to train the model using a hybrid loss function, consisting of evidential learning loss and two regularization terms. Here, we demonstrate the effectiveness of two proposed regularization terms. Table 4 shows the quantitative results of models trained using different loss functions. The evidential learning loss is the original loss without any regularization (loss-1). The regular-

ization term based on predictions minimizes the evidence v and α on erroneous regions (loss-2). The regularization term based on disparity agreement smooths the evidence on the regions with disparity agreement (loss-3). The combination of these two regularization terms ensures the constraints on all pixels (loss-4) and achieves the highest performance. The learning curves of loss functions are shown in Figure 9. Loss_total is the whole loss function. Loss_data is the evidential deep learning loss, which is calculated using annotated pixels. Loss_pred regularizes the uncertainty parameters, especially the annotated pixels. Loss_smooth leverages the information from annotated pixels and penalizes surrounding unannotated pixels. The loss_pred and loss_smooth make it possible to constrain all positions in the stereo image, obtaining better uncertainty estimation results.

5. Conclusion

In this paper, we introduce a deep evidential learning-based approach for stereo matching and uncertainty estimation. By considering learning as an evidence acquisition process and estimating the parameters of NIG distribution, our proposed approach can obtain disparity, aleatoric and epistemic uncertainty in an end-to-end way. It effectively reduces the time and computation consumption when estimating uncertainty. The proposed two loss terms propagate the supervisory signal so that the network can be trained well even with only sparse disparity annotations. The current approach also has two main weaknesses. Firstly, ground truth disparity annotations are still needed, which is not conducive to training on large-scale datasets. We will study a completely self-supervised uncertainty estimation method. Secondly, there is currently no demonstrable mechanism for using uncertainty to improve stereo matching results. In particular, epistemic uncer-

tainty can capture out-of-distribution data, which has the potential to solve domain adaptation problems. In the future, we will concentrate on these two problems.

Acknowledgment

This work was supported by the Beijing Natural Science Foundation(4202039), NSFC No. 61772057 and No. 61901436. The support funding was also from Key Research Program of the Chinese Academy of Sciences, Grant NO. XDPB22, State Key Lab. of Software Development Environment and Jiangxi Research Institute of Beihang University.

References

[1] H. Liu, X. Tang, S. Shen, Depth-map completion for large indoor scene reconstruction, Pattern Recognition 99 (2020) 107112.

[2] R. Munoz-Salinas, M. J. Marin-Jimenez, R. Medina-Carnicer, Spm-slam: Simultaneous localization and mapping with squared planar markers, Pattern Recognition 86 (2019) 156–171.

[3] H. Wang, T. Xu, J. Guo, Z. Rao, G. Shi, Withdrawn: Incremental subspace and probability mask constrained tracking in smart and autonomous systems, Pattern Recognition 72 (2017) 473–483.

[4] J. R. Chang, Y. S. Chen, Pyramid stereo matching network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5410–5418.

[5] F. Zhang, V. Prisacariu, R. Yang, P. H. Torr, Ga-net: Guided aggregation net for end-to-end stereo matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 185–194.

[6] H. Xu, J. Zhang, AANet: Adaptive aggregation network for efficient stereo matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 1959–1968.

[7] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter?, Structural safety 31 (2) (2009) 105–112.

[8] X. Hu, P. Mordohai, A quantitative evaluation of confidence measures for stereo vision, IEEE Transactions on Pattern Analysis and Machine intelligence 34 (11) (2012) 2121–2133.

[9] M. Poggi, F. Tosi, S. Mattoccia, Quantitative evaluation of confidence measures in a machine learning world, in: IEEE International Conference on Computer Vision, 2017, pp. 5228–5237.

[10] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, 2017, pp. 5574–5584.

[11] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: International Conference on Machine Learning, 2016, pp. 1050–1059.

[12] J. Postels, F. Ferroni, H. Coskun, N. Navab, F. Tombari, Sampling-free epistemic uncertainty estimation using approximated variance propagation, in: IEEE International Conference on Computer Vision, 2019, pp. 2931–2940.

[13] A. Amini, W. Schwarting, A. Soleimany, D. Rus, Deep evidential regression, in: Advances in Neural Information Processing Systems, 2020.

[14] A. Malinin, M. Gales, Predictive uncertainty estimation via prior networks, in: Advances in Neural Information Processing Systems, 2018, pp. 7047–7058.

[15] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Advances in Neural Information Processing Systems, 2018, pp. 3179–3189.

[16] D. Scharstein, R. Szeliski, A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, International Journal of Computer Vision 47 (1-3) (2002) 7–42.

[17] Z. Jie, P. Wang, Y. Ling, B. Zhao, Y. Wei, J. Feng, W. Liu, Left-right comparative recurrent model for stereo matching, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3838–3846.

[18] F. Cheng, X. He, H. Zhang, Learning to refine depth for robust stereo estimation, Pattern Recognition 74 (2018) 122–133.

[19] B.-S. Shin, D. Caudillo, R. Klette, Evaluation of two stereo matchers on long real-world video sequences, Pattern Recognition 48 (4) (2015) 1113–1124.

[20] J. Pang, W. Sun, J. Ren, C. Yang, Q. Yan, Cascade residual learning: A two-stage convolutional neural network for stereo matching, in: International Conference on Computer Vision-Workshop on Geometry Meets Deep Learning, 2017, pp. 878–886.

[21] X. Cheng, P. Wang, R. Yang, Learning depth with convolutional spatial propagation network, IEEE Transactions on Pattern Analysis and Machine Intelligence (2019).

[22] C. Wang, X. Bai, X. Wang, X. Liu, J. Zhou, X. Wu, H. Li, D. Tao, Self-supervised multiscale adversarial regression network for stereo disparity estimation, IEEE Transactions on Cybernetics (2020).

[23] F. Zhang, X. Qi, R. Yang, V. Prisacariu, B. Wah, P. Torr, Domain-invariant stereo matching networks, in: European Conference on Computer Vision, Springer, 2020, pp. 420–439.

[24] O. Choi, H. S. Chang, Yet another cost aggregation over models, IEEE Transactions on Image Processing 25 (11) (2016) 5397–5410.

[25] A. Shaked, L. Wolf, Improved stereo matching with constant highway networks and reflective confidence learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4641–4650.

[26] S. Kim, D. Min, S. Kim, K. Sohn, Unified confidence estimation networks for robust stereo matching, IEEE Transactions on Image Processing 28 (3) (2018) 1299–1313.

[27] S. Kim, D. Min, S. Kim, K. Sohn, Adversarial confidence estimation networks for robust stereo matching, IEEE Transactions on Intelligent Transportation Systems (2020).

[28] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems, 2017, pp. 6402–6413.

[29] E. D. Carvalho, R. Clark, A. Nicastro, P. H. Kelly, Scalable uncertainty for computer vision with functional variational inference, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 12003–12013.

[30] Y. Hu, W. Zhen, S. Scherer, Deep-learning assisted high-resolution binocular stereo depth reconstruction, in: IEEE International Conference on Robotics and Automation, 2020, pp. 8637–8643.

[31] M. Poggi, F. Aleotti, F. Tosi, G. Zaccaroni, S. Mattoccia, Self-adapting confidence estimation for stereo, in: European Conference on Computer Vision, 2020, pp. 715–733.

[32] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, T. Brox, A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4040–4048.

[33] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The KITTI vision benchmark suite, in: IEEE Computer Vision and Pattern Recognition, 2012, pp. 3354–3361.

[34] M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3061–3070.

[35] W. Bao, W. Wang, Y. Xu, Y. Guo, S. Hong, X. Zhang, Instereo2k: a large real dataset for stereo matching in indoor scenes, Science China Information Sciences 63 (11) (2020) 1–11.

[36] M. S. K. Gul, M. Bätz, J. Keinert, Pixel-wise confidences for stereo disparities using recurrent neural networks., in: British Machine Vision Conference, 2019, p. 23.