Predicting Disparity Distributions

Gustav Häger, Mikael Persson and Michael Felsberg

Häger, G., Persson, M., Felsberg, M., (2021), Predicting Disparity Distributions, *2021 IEEE International Conference on Robotics and Automation (ICRA)*.
https://doi.org/10.1109/ICRA48506.2021.9561617

Predicting Disparity Distributions

Gustav Häger[1], Mikael Persson[1], Michael Felsberg[1]

***Abstract*— We investigate a novel deep-learning-based approach to estimate uncertainty in stereo disparity prediction networks. Current state-of-the-art methods often formulate disparity prediction as a regression problem with a single scalar output in each pixel. This can be problematic in practical applications as in many cases there might not exist a single well defined disparity, for example in cases of occlusions or at depth-boundaries. While current neural-network-based disparity estimation approaches obtain good performance on benchmarks, the disparity prediction is treated as a black box at inference time. In this paper we show that by formulating the learning problem as a regression with a distribution target, we obtain a robust estimate of the uncertainty in each pixel, while maintaining the performance of the original method. The proposed method is evaluated both on a large-scale standard benchmark, as well on our own data. We also show that the uncertainty estimate significantly improves by maximizing the uncertainty in those pixels that have no well defined disparity during learning.**

I. Introduction

Many autonomous system tasks such as motion-planning and navigation require accurate estimates of the 3D environment geometry. One approach to obtaining these 3D estimates is using photometric stereo cameras. A real-world system is likely to combine information from multiple modalities, necessitating accurate estimates of the accuracy in each method. Current state-of-the-art approaches to photometric stereo are often based on using convolutional neural networks (CNNs) to obtain disparity measurements, but ignore the issue of obtaining uncertainty estimates.

Current CNN-based stereo methods formulate the problem of disparity estimation as learning a mapping Z between a reference I_l and a target image I_r to an estimate of the disparity d in each pixel of the reference image. The output is typically a scalar-valued disparity estimate $\hat{d}$ in each pixel. Most approaches treat the estimator Z as a black-box during inference, making it challenging to obtain robust estimates of the uncertainties in each pixel.

Obtaining uncertainty information from CNNs is an active research area. Approaches can be separated into those estimating uncertainties for model choices (epistemic uncertainty) or for the predictions (aleatoric uncertainty) [1].

Estimating uncertainty in stereo methods is often done as a separate post-processing step after the disparity prediction has been generated. In the closely related field of optical flow it has been proposed to obtain uncertainties jointly with the predictions by representing the predictions as parameterized distributions instead. This is done either using ensembles or by direct prediction of the parameters [2].

In this paper we address this problem of jointly predicting disparity and uncertainty. We employ an information theoretic approach for obtaining the uncertainties. Instead of only predicting a single disparity $\hat{d}$ in each pixel, we instead predict a distribution $\hat{P}_d$ in each pixel. By viewing the output of the network as a density estimator, uncertainty information is represented in the shape of the output distribution.

Using a density estimate as the representation of the network output has the advantage that we can represent situations such occlusions as a distribution with maximal uncertainty. The traditional approach of directly producing a prediction $\hat{d}$ in each pixel, has the advantage that the L_1 error can be minimized directly. We show that the advantages of both approaches can be kept by minimizing the Wasserstein distance between P_d and $\hat{P}_d$.

II. Related work

Many classical rectified stereo methods follow a similar structure: (a) feature representations of both images are extracted, (b) matching costs are computed along the epipolar lines, resulting in a 3D volume, (c) the volume is regularized, and (d) a single disparity is selected in each pixel, often by solving an optimization problem [3]. An example of such a 3D volume is shown in figure 1. Early approaches utilizing learning mainly applied to parts of the described pipeline, such as learning a scoring function [4].

Recent works typically learn feature extraction, matching scores, and regularization jointly by formulating a regression directly from the input images to the disparities. Such methods often include a 3D matching volume as an intermediate step. This volume is processed using 3D convolutions and the disparity estimate is obtained through a differentiable approximation of the argmax operation [5], [6], [7], [8]. This differentiable argmax, often called the soft-argmax, allows direct minimization of the geometric error of the predictions. In many cases a smooth L_1 loss [6], [5], [8], [9] is used.

The soft-argmax operation is a maximum a posteriori estimator, in cases of unimodal and symmetric distributions

[1]firstname.lastname@liu.se

This work was supported by the Wallenberg AI, Autonomus Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Computational resources where provided by the Swedish National Infrastructure for Computing (SNIC) at C3SE, as well as Daimler Ag, and Singulareye.

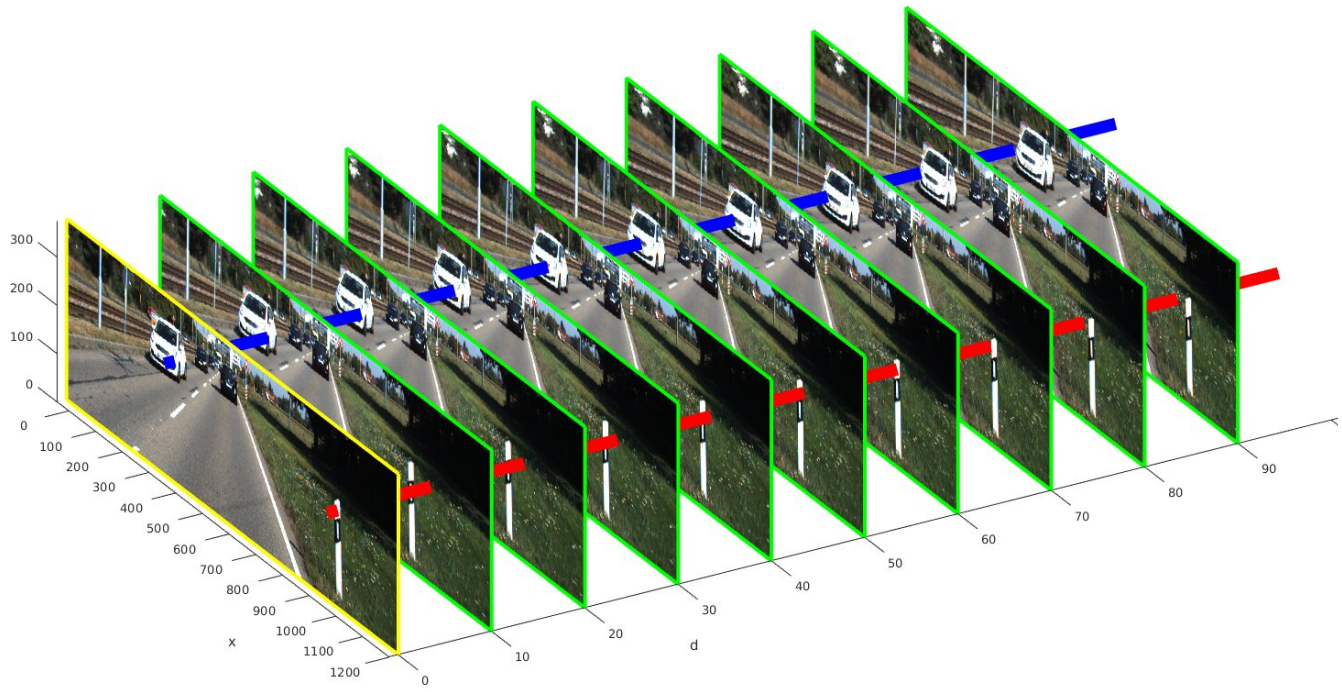

Fig. 1. Visualization of the matching step of a general stereo algorithm. Yellow and green borders denote the reference and target image respectively. Each step in the d-direction corresponds to one step to the left in the target image. The blue and red lines correspond to epipolar lines from the reference image. At the true disparity, the line intersects the target image in the same 3D point as in the reference image.

[7]. For distributions where these assumptions do not hold, that is, multimodal and/or nonsymmetric distributions, it can be shown to be a biased estimator [7]. It has been argued to not be a problem in practice, as the network is forced to learn to compensate for this bias when trained through the estimator [5]. However, we argue that this is not true in general, as cases where the disparity is ambiguous will result in the soft-argmax operation blending multiple modes. Furthermore it has been noted that by back-propagating through a softmax-layer, as in the soft-argmax operation, the output distribution will be highly overconfident [10].

It has been proposed to reduce this bias by instead predicting a distribution, and minimizing the cross-entropy or a classification loss [5]. This was taken further by [7] who attempted to further reduce the bias minimizing the cross-entropy with respect to a Laplace distribution, for a small region centered on the max of $\hat{P}_d$. These methods minimize the cross-entropy error, rather than the L_1 error.

The task of uncertainty estimation for general stereo methods has been investigated extensively in the literature, in particular [11] proposes a taxonomy of uncertainty measures for disparity, while [12] proposes an updated version accounting for early deep-learning based approaches to disparity estimation. A common theme with many of these methods is that they operate on the range of scores generated by a stereo method [11], [12], in our notation this is the estimated disparity distribution $\hat{P}_d$. They separate these methods into local, global (for a single $\hat{P}_d$), left-right consistency checking, disparity map, reference image features, or learning based approaches [12]. Common to the discussed approaches, is that they are expected to be applied after the disparity is already calculated, and apply equally well to any stereo method [12]. In our setting we wish to estimate the disparity and an uncertainty jointly, rather than as a post-processing step, without a separate network branch for uncertainty estimates.

Our approach for uncertainty estimation is closest to that of Ilg et al. [2] who obtained uncertainties and optical flow predictions jointly parametric distributions. The authors evaluate several approaches for obtaining the parameters of this distribution, such as Monte-Carlo dropout followed by fitting the parameters of the distribution, to direct prediction of parameters in the network itself.

However unlike [2] and [13] we do not use a parameterized model for our distribution. Instead the parameterization is learned as a part of the neural network. By doing this we avoid implicitly enforcing a particular shape of our predicted distributions. Earlier work refer to this type of learned distribution as mixture-density networks [14], [15]. An advantage of this approach is that the network is capable of outputting a distribution of arbitrary shapes, unlike when a parameterized model is used.

Creation of a large scale dataset with annotated disparities is a challenging problem, as individually labeling each individual pixel with a correct disparity is extremely labor-intensive. Instead, synthetic data sets such as [16] are typically used [6], [5], [7]. The synthetic images is rendered along with the corresponding ground-truth by utilizing the depth-buffer of the renderer. This means that ground-truth is available for all pixels in an image, even if they correspond to 3D points visible in only one of the images. It has been shown that pre-training on this synthetic data before fine-tuning on smaller amounts of real-data works well in practice [6], [5], [7], [8], [9].

III. Predicting disparity

We consider the problem of disparity estimation from rectified stereo images. The goal is estimating the offset for a projected 3D point in the reference image I_l to its corresponding position in the target image I_r. In our case we use the left image as I_l, and the right image as I_r. A 3D point projected to the position (x, y) in the left image, will assuming it is visible in both images, be projected to the position $(x - d, y)$ in I_r. Recovering the depth r of a pixel can be done by computing $r = \frac{fb}{d}$, where f is the camera focal-length and b the baseline between the camera centers. Note that it is possible for a 3D point to be visible in only one of I_l or I_r. In such cases the disparity can be considered to either not exist at all, or as a function of the inverse depth.

For a single pixel, the range of possible disparities is $d \in [0, D]$ where D is some maximum disparity. By considering this range for each (x, y) position in I_l, a 3D volume with indices (x, y, d) is obtained. A visualization of this concept can be seen in figure 1 where the blue and red lines correspond to epipolar lines for two different pixels.

Many CNN-based stereo methods construct the estimation network Z in a way similar to that of a plane sweeping stereo method: (a) 2D features are extracted from I_l and I_r, (b) copies of I_l are concatenated with shifted copies of I_r, and (c) the volume is processed using 3D convolutions. This means that a step along the disparity dimension corresponds to a step along the epipolar line for that pixel in the corresponding image. In such frameworks the matching scores from a classical stereo method are defined implicitly, and the optimization procedure is replaced with the soft-argmax. This enables the parameters of Z to be learned by minimizing the L^1 error between the ground-truth disparity and the prediction $\hat{d}$.

As the soft-argmax operation does not have any learnable parameters, we define Z to be the network without the soft-argmax applied. We use the network from [6] for all our experiments unless otherwise noted. This allows us to view Z as assigning a matching score to every possible disparity for each pixel. For brevity we omit the pixel indices, writing the score assigned to disparity d for any pixel as: $Z(d)$. In this notation the soft-argmax operation is defined as:

$$\hat{d} = \sum_{n}^{D} \frac{n e^{Z(n)}}{\sum_{m}^{D} e^{Z(m)}} \ . \tag{1}$$

This operation can be viewed as a combination of a softmax operation along with a mode-estimation. The softmax operation normalizes the output of Z to be a distribution, while suppressing weaker modes:

$$\hat{P}_d(d) = \frac{e^{Z(d)}}{\sum_{m}^{D} e^{Z(m)}} \ . \tag{2}$$

This results in a distribution $\hat{P}_d$ for each pixel, where the mode estimation step can be expressed in terms of this distribution as:

$$\hat{d} = \sum_{n}^{D} n \hat{P}_d(n) \ . \tag{3}$$

However this mode-estimator assumes the distribution is unimodal. As an example of this not being the case, the red line in figure 1 can be considered to lie both on and beside the white pole, generating two possible disparities. In such a case the soft-argmax will blend both-modes.

IV. Learning to predict disparity distributions

In order to obtain uncertainties along with the disparities, we propose to reformulate the learning of Z such that it directly estimates distributions over possible d. In this framework we can interpret the shape of $\hat{P}_d$ as an uncertainty in a principled way. We denote the predicted distribution for a single pixel as $\hat{P}_d$. Unlike in [2] we do not impose a parameterized representation for $\hat{P}_d$. Instead we consider the output of Z as density estimate for $\hat{P}_d$. This interpretation is supported by the fact that every position in the output volume corresponds directly to a certain disparity for a certain pixel.

A. Learning the distribution

We alter the baseline network [6] to output a density estimate by removing the softmax function. In order to ensure the output is positive we apply a softplus function to each of the outputs, followed by normalization. This corresponds to replacing the softmax function with an activation function written as:

$$\hat{P}_d(d) = \frac{\rho(Z(d))}{\sum_{n}^{D} \rho(Z(n))} \ , \tag{4}$$

where ρ is the softplus operation defined as:

$$\rho(x) = \log(1 + e^x) \ . \tag{5}$$

Using this activation function the output of our network can be interpreted as a distribution over possible disparities in each pixel. While this is also true for the softmax function, it is highly nonlinear and best suited to minimizing a cross-entropy loss over categorical distributions [17]. The softplus function instead, is approximately linear, while maintaining positive output.

What remains is then to construct a model distribution for the ground-truth. Most methods implicitly assume that the ground-truth has a single exact value, implicitly modeling the density as being concentrated at the given disparity. In a distribution framework this can be expressed as a dirac impulse at the ground-truth disparity.

If we model both the network output and the ground-truth as Dirac impulses δ_d, and $\delta_{\hat{d}}$, the absolute error $|d - \hat{d}|$ can be minimized during training by means of the L_1 norm of the integral difference of the distributions. This is similar to using the MAP approach, if the variance is ignored. If we define $F_d(t)$, and $\hat{F}_d(t)$ as the cumulative distribution functions of P_d and $\hat{P}_d$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{t} |P_d(\tau) - \hat{P}_d(\tau)| \ \mathrm{d}\tau \ \mathrm{d}t \ = \int_{-\infty}^{\infty} |F(t-d) - F(t-\hat{d})| \ \mathrm{d}t, \tag{6}$$

in each pixel. Here we used that F_d is the primitive function of P_d, and assume identically shaped distributions.

This is the case when training using ground-truth of some known variance, but only producing a point-estimate as prediction. By using the fundamental theorem of calculus we can rewrite this as:

$$\int_{-\infty}^{\infty} | \int_{\hat{d}}^{d} P(t-\tau) \ \mathrm{d}\tau | \ \mathrm{d}t = | \int_{\hat{d}}^{d} 1 \ \mathrm{d}\tau | = |d - \hat{d}| \ . \quad (7)$$

Minimizing such an error function, rather than the more common cross entropy, maintains the geometric property that predictions far away from the ground-truth should give a larger error than those that are close to the ground-truth. Furthermore, if the distributions are non-overlapping the gradient of cross-entropy loss vanishes.

As our network output representations is a density estimate of $\hat{P}_d$, rather than a parameterization with known shape, we need a way to compare the prediction with a general density of the ground-truth P_d, meaning that (6) needs to be generalized to handle different distributions for P_d and $\hat{P}_d$. If we consider the total probability-mass transport required to transform $\hat{P}_d$ into P_d as the distance between the distributions, we can maintain the geometric interpretation of the original problem, while still representing the prediction and ground-truth as general distributions. This metric is usually known as the Wasserstein distance, when using the L^1 distance, the Wasserstein-1 or W^1 distance:

$$W^1(P_d, P_{\hat{d}}) = \inf_{\gamma \in \Gamma(P, P_{\hat{d}})} \int_{-\infty}^{\infty} |d - \hat{d}| \, d\gamma(d, \hat{d},) \ , \quad (8)$$

where the set of all plans for turning $\hat{P}_d$ into P_d is denoted as $\Gamma(P_d, \hat{P}_d)$. The infimum is taken over all possible plans in Γ for turning $\hat{P}_d$ into P. This metric is well-defined even for distributions such as δ, or general non-overlapping ones.

For general multi-variate distributions, computing $W^1(P_d, \hat{P}_d)$ requires the solution of the optimal transport problem due to the infimum in the definition. However in the case of one-dimensional distributions, the optimization problem has a simple closed-form solution in terms of F_d and $\hat{F}_d$. This solution is well known in statistics literature [18], [19], for a simple proof we refer to [18]. As reproducing the proof itself does not offer further insights into the task of disparity estimation we just repeat the result here:

$$W^1(P_d, \hat{P}_d) = \int_0^D |F_d(t) - \hat{F}_d(t)| \, dt \ . \quad (9)$$

Only the interval $[0, D]$ is of interest for both distributions, as we know that the probability is zero outside this range for both P_d and $\hat{P}_d$. Applying back-propagation through this expression allows us to find the parameters of Z for a general input and output distribution by minimizing $W^1(P, \hat{P}_d)$ in each pixel. For numerical reasons it is useful to apply a small smoothing to the ground-truth distributions.

B. Uncertainties from entropy

By formulating Z to output a non-parametric density estimate we avoid restricting the output distribution to a (usually uni-modal) family of parametric distributions. In this way the output of the network can better reflect the possible multi-modality at depth-discontinuities. Moreover, it allows the output of our network to be interpreted as a matching score for each possible disparity.

At depth-discontinuities the true disparity is ambiguous as a pixel can intersect both foreground and background objects. In such situations the variance will be dependent on the distance between the two modes. The consequence of this is that the uncertainty estimate is more dependent on the distance between foreground and background than the fact that there exists multiple possible disparities.

A robust uncertainty measure should be guaranteed to increase with he number of peaks, regardless of their positions. For this reason we utilize the Shannon entropy [20] of the output distribution as our uncertainty estimator:

$$H = -\sum_{n}^{D} \hat{P}_d(n) \log(\hat{P}_d(n)) \ . \quad (10)$$

Unlike variance, the entropy does not measure the distance to some hypothetical mode of the predictions, instead it measures the ambiguity of the prediction. The entropy is minimized in the case of $P_{\hat{d}}$ being a point-mass, and maximized for the uniform distribution. The interpretation being that a point-mass has only a single possible value, and the uniform distribution considers every disparity as equally likely. Unlike the variance, the entropy is guaranteed to increase as more peaks are added to $P_{\hat{d}}$.

This also allows us to express the absence of disparity in particular pixels, such as in the case of occlusions or due to the true disparity being larger than the maximum D, by setting the label during training to the uniform distribution.

V. Experiments

We evaluate our proposed approach comprehensively on the scene flow dataset [16], a novel dataset. The evaluations utilize two separate criteria, the endpoint-error (EPE) and sparsification curves. The EPE measures the performance for the disparity prediction task, while the sparsification curves give a measure of how well the predicted uncertainty corresponds to pixels with high errors.

Our network is based on PSMnet [6]. We compare several methods for uncertainty estimation: A naive approach obtaining uncertainty from the output of the softmax layer in PSMnet, an ensemble of PSMnets, as well two different approaches for training distribution prediction networks. Note that our main goal is not to improve the EPE, but rather to show the viability of our method for estimating the uncertainty of each prediction. In order to compare with methods other than our own, we also train from scratch an instance of the DeepPruner [9] network. As this network does not produce uncertainties it is not included in the uncertainty evaluation.

In order to validate our approach we first perform an ablation study using the synthetic scene-flow benchmark. In this setting we have ground-truth annotations for all pixels, including those that are occluded, as the ground-truth is generated directly from depth buffer of the renderer. The

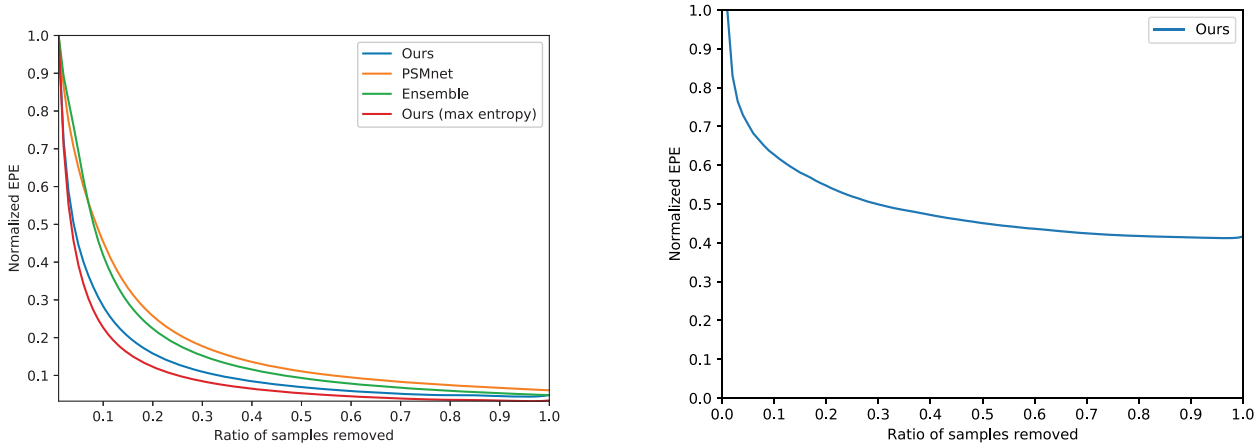

Fig. 2. The sparsification plot is constructed by removing a ratio of the predictions, sorted by uncertainty. The x-axis corresponds to the portion of removed samples, and the y-axis the mean EPE for the remaining predictions. As predictions are removed in order of decreasing uncertainty, an uncertainty measure that accurately predicts a high error will decrease more sharply as the unreliable measurements are removed. The left image shows a baseline comparison on the Freiburg dataset, the right image shows the performance of our method on the IMO dataset.

TABLE I

BASELINE COMPARISON

Method	EPE	EPE $< D$	$e < 1$	$e < 2$	$e < 3$
PSMNet [6]	2.70	1.37	83%	91%	94%
DeepPruner [9]	2.55	1.40	80%	89%	91%
Ensemble	3.36	1.78	83%	89%	94%
Ours	2.70	1.41	84%	91%	93%
Ours (max-entropy)	2.70	1.41	84%	91%	93%

TABLE II

COMPARISON ON OUR OWN DATASET.

Method	EPE	$e < 1$	$e < 2$	$e < 3$
DeepPruner [9]	2.70	46%	74%	89%
PSMNet [6]	1.30	52%	76%	87%
Ours	0.87	67%	93%	98%
Ours (max entropy)	1.30	48%	77%	89%

methods based on PSMNet are trained as proposed by the authors [6]. That is, we calculate the loss during training only for those pixels where $d < D$. The learning rate is set to 10^{-4} following [6], using randomly cropped (256, 512) sized patches from the input images.

In the evaluation we report both the EPE including only pixels with disparity less than the maximum, as this follows the protocol of many other authors [6], [5], [7], [9]. In a realistic setting there is no way to control what the maximum disparity in the input image is, so we also include the EPE calculated over all pixels, including those with ground-truth values greater than the maximum disparity.

Our baseline approach is an unmodified PSMNet [6], trained on the flying-things subset of the Freiburg dataset, as described by the authors [6]. We modify this network to use 4 as activation. This means that the output of our network is a density estimate in each pixel for the disparity range $d \in [0, D]$. The parameters of Z are learned by minimizing the W^1 error with respect to a ground-truth distribution. In order to avoid numerical issues, we apply a small blurring to the ground-truth distributions. We obtain the uncertainty as the entropy of the softmax distribution, and the predicted distribution respectively. Point estimates for our method are obtained by calculating the mode of the predicted distribution.

We also investigate the performance impact of training the network to predict a maximum-entropy distribution for those pixels where the true disparity match is occluded. We find occluded pixels by a left-right consistency scheme. The network trained using this approach is noted as max-entropy in the tables.

Finally we include an ensemble of PSMNets, where the predicted distribution is obtained by fitting a Gaussian distribution to the predictions. The uncertainty is obtained as the variance of this distribution, and the mean used as the point-estimate. The ensemble consists of five networks trained using the same parameters, on non-overlapping subsets of the training data.

The results of our comparison is shown in table I. While DeepPruner [9] obtains the lowest average EPE for the full set of disparities, it clearly has a higher minimum error seen from the comparatively lower number of pixels with an EPE less than 1. The PSMnet method obtains the lowest EPE when considering only pixels with less disparity than the maximum value, for the more realistic setting where all pixels are included it is tied in second place with our method trained using the Wasserstein loss. Finally our proposed entropy maximization method obtains comparable performance with respect to EPE, but as shown in figure 2 significantly outperforms the other methods in uncertainty estimation.

The uncertainty is evaluated in terms of sparsification curves as done by [2]. A sparsification curve plots the average EPE over all pixels, where increasing numbers of data points are removed based on the predicted uncertainty. If the uncertainty prediction is accurate, the average error will be reduced by removing the most uncertain predictions.

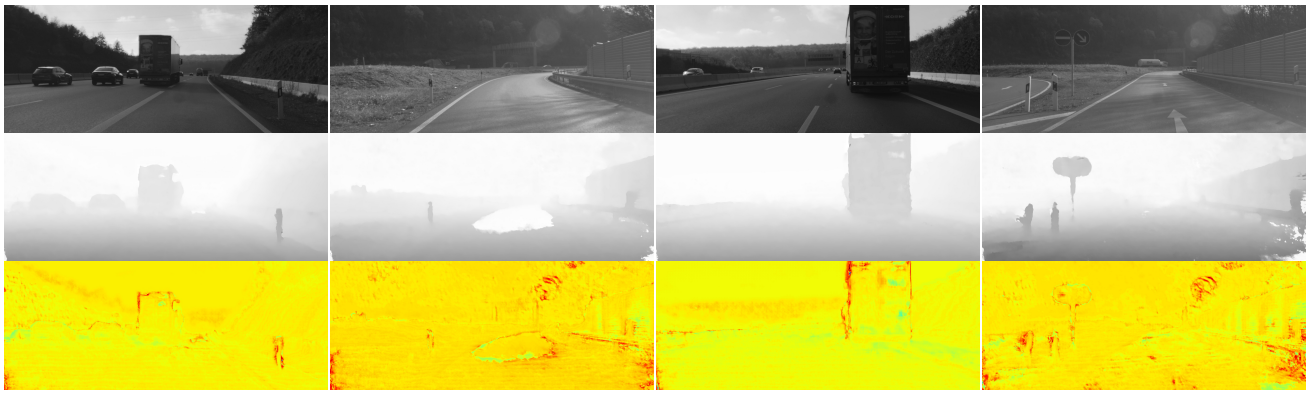

Fig. 3. Qualitative results for our entropy maximization method on our own dataset. From top to bottom: The left input image, the predicted disparity and the predicted uncertainty. The uncertainty is clearly higher in regions at object boundaries, as well as at regions with smaller objects.

If the uncertainty estimation is poor, it will not correctly predict pixels with high errors.

The naive baseline of taking the entropy of the softmax distribution has the least reliable uncertainty estimates. This is likely due to the softmax function suppressing minor modes of the output. The ensemble approach performs slightly better at uncertainty estimation, but is hampered both by the fairly small number of components, and its uncertainty estimate being a variance of unimodal distribution as discussed in section IV. The second best performing approach is our proposed distribution prediction method, when trained using all pixels in the ground-truth, without accounting for visibility. The best uncertainty estimates are obtained by also training the method to maximize the entropy for occluded pixels. This method significantly improves the uncertainty estimate, with a tradeoff of slightly worse EPE.

In order to show the performance of our approach on large-scale real data, we use the IMO dataset. This dataset has 5000 stereo images collected from a moving car. We obtain ground-truth disparities by utilizing a visual-odometry system based on [21], where the estimated 3D coordinates are fine-tuned using a bundle-adjustment step minimizing the re-projection error. Only points with a re-projection error of less than one pixel in four consecutive frames are considered robust enough to be included in the ground-truth. This approach allows us to obtain sparse, but accurate ground-truth for a large number of real-world stereo image pairs. The disparity is calculated from the inverse of the depth for each of the selected pixels. This dataset is available for download[1].

For the real-data experiment all methods are first pre-trained on the Freiburg dataset for 10 epochs, followed by 10 epochs of fine-tuning on the real data. The learning rate for fine-tuning was the same as when training the initial model. Here the best method is our proposed distribution prediction approach, achieving an average EPE of less than 1 pixel. The entropy maximization approach performs slightly worse, being tied with PSMNet at 1.30 EPE, while DeepPruner error is far higher at 2.7 EPE. This difference in EPE seems to mainly occur at higher errors, as the number of pixels below each threshold in II is similar.

[1]https://www.cvl.isy.liu.se/research/datasets/imo-dataset/

Example disparity and uncertainty predictions, along with the corresponding input images are demonstrated in Figure 3. In the first frame the uncertainty is very high at the road-marker, as the network cannot determine if a pixel belongs to the marker or the background. The second image demonstrates a failure case of our method, where the disparity prediction completely fails on the homogeneous road surface, while the uncertainty is slightly higher it fails to mark the entire region as unreliable. The third image demonstrates how the uncertainty prediction is accurately marking areas that are only visible in one of the input images, such as the side of the truck. In the second and last examples the uncertainty estimation correctly identifies the lens flare as a problematic region. In all images the uncertainty is high in areas with ambiguous disparities, such as object boundaries, particularly around smaller objects.

Overall our proposed uncertainty estimation method succeeds at accurately predicting problematic pixels, a quantitative measure can be seen in the sparsification curve for the IMO dataset in figure 2 (right). Here the EPE improvement saturates when approximately 20% of the points are removed. This is likely a consequence of the ground-truth generation, as it effectively reduces the number of difficult pixels in the dataset, particularly occluded pixels are completely absent from the ground-truth data.

In this dataset the EPE improvement saturates at approximately 20% of points removed. This is likely due the ground-truth generation process, that unlike for the synthetic data does not have ground-truth disparities for occluded pixels. This means that even when the uncertainty estimate correctly detects such regions, it will not improve the average EPE for pixels with ground-truth.

VI. Conclusions

We have proposed an approach for jointly predicting disparity and uncertainties for a stereo network. Our method achieves accurate disparity predictions, as well as robust uncertainty estimates on both simulated and real data. The uncertainty estimate is obtained jointly with the prediction of disparity, without additional post-processing. The uncertainty predictions can be further improved by adjusting the learning to maximize uncertainty in regions occluded regions. This

improvement is obtained without significant loss of performance in the disparity predictions.

REFERENCES

[1] A. Kendall and 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.

[2] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 652–667.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International journal of computer vision*, vol. 47, no. 1-3, pp. 7–42, 2002.

[4] J. Žbontar and Y. LeCun, "Stereo matching by training a convolutional neural network to compare image patches," *The journal of machine learning research*, vol. 17, no. 1, pp. 2287–2318, 2016.

[5] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 66–75.

[6] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[7] S. Tulyakov, A. Ivanov, and F. Fleuret, "Practical deep stereo (pds): Toward applications-friendly deep stereo matching," in *Advances in Neural Information Processing Systems*, 2018, pp. 5871–5881.

[8] R. Chabra, J. Straub, C. Sweeney, R. Newcombe, and H. Fuchs, "Stereodrnet: Dilated residual stereonet," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 786–11 795.

[9] S. Duggal, S. Wang, W.-C. Ma, R. Hu, and R. Urtasun, "Deeppruner: Learning efficient stereo matching via differentiable patchmatch," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4384–4393.

[10] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[11] X. Hu and P. Mordohai, "A quantitative evaluation of confidence measures for stereo vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2121–2133, 2012.

[12] M. Poggi, F. Tosi, and S. Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[13] J. Gast and S. Roth, "Lightweight probabilistic deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3369–3378.

[14] C. M. Bishop, "Mixture density networks," 1994.

[15] O. Makansi, E. Ilg, O. Cicek, and T. Brox, "Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7144–7153.

[16] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, arXiv:1512.02134.

[17] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[18] A. Ramdas, N. G. Trillos, and M. Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.

[19] O. Thas, *Comparing distributions*. Springer, 2010.

[20] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[21] M. Persson, T. Piccini, M. Felsberg, and R. Mester, "Robust stereo visual odometry from monocular techniques," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2015, pp. 686–691.