

Learning the Distribution of Errors in Stereo Matching for Joint Disparity and Uncertainty Estimation

Liyan Chen Weihan Wang Philippos Mordohai
 Stevens Institute of Technology

Abstract

We present a new loss function for joint disparity and uncertainty estimation in deep stereo matching. Our work is motivated by the need for precise uncertainty estimates and the observation that multi-task learning often leads to improved performance in all tasks. We show that this can be achieved by requiring the distribution of uncertainty to match the distribution of disparity errors via a KL divergence term in the network’s loss function. A differentiable soft-histogramming technique is used to approximate the distributions so that they can be used in the loss. We experimentally assess the effectiveness of our approach and observe significant improvements in both disparity and uncertainty prediction on large datasets. Our code is available at <https://github.com/lly00412/SEDNet.git>.

1. Introduction

Many computer vision problems can be formulated as estimation tasks. Considering, however, that even high-performing estimators are not error-free, associating confidence or uncertainty with their estimates is of great importance, particularly in critical applications. In this paper, we focus on disparity estimation via stereo matching, but we are confident that our approach is applicable to other pixel-wise regression tasks after minor modifications.

We distinguish between *confidence* and *uncertainty*: the former refers to a probability or likelihood of correctness, while the latter is related to the magnitude of the expected error of an estimate. Confidence can be used to reject estimates that are suspected to be incorrect, or to rank them from most to least reliable. We argue that uncertainty is more valuable because it can also be used for fusing multiple observations, e.g. in a Kalman filtering framework. Most research has focused on confidence estimation for stereo matching [12, 30]. Moreover, most methods estimate confidence for pre-computed disparities that are not further improved. Joint estimation of disparity and confidence, which benefits both due to multi-task learning, is addressed infrequently [19, 20, 26, 34].

Our work is partially inspired by the joint treatment of epistemic and aleatoric uncertainty by Kendall and Gal [14], who propose novel loss functions that give rise to uncertainty estimates in pixel-wise vision tasks. Results on semantic segmentation and single-image depth estimation demonstrate how the primary task benefits from simultaneous uncertainty estimation. Kendall and Gal argue that “in many big data regimes (such as the ones common to deep learning with image data), it is most effective to model aleatoric uncertainty,” while epistemic uncertainty can be reduced when large amounts of data are available. Here, we restrict our attention to aleatoric uncertainty.

Our motivation is that ideally we should be able to predict the magnitude of the estimator’s error at each pixel. Of course, this is unrealistic, since if it was possible, we could drive all errors down to zero. A feasible objective is to train an *uncertainty estimator whose outputs follow the same distribution as the true errors of the disparity estimator*.

In this paper, we present an implementation of this concept via a deep network that jointly estimates disparity and its uncertainty from a pair of rectified images. We named the network **SEDNet**, for *Stereo Error Distribution Network*. SEDNet includes a novel, lightweight uncertainty estimation subnetwork that predicts the aleatoric uncertainty

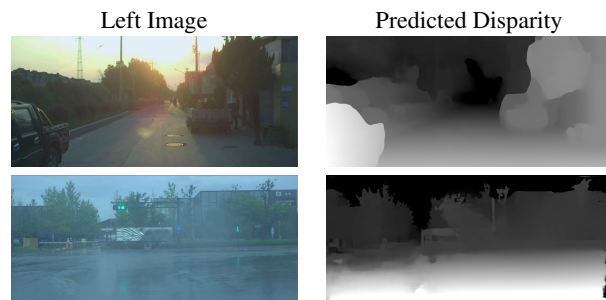


Figure 1. Examples of left images and predicted disparity maps by SEDNet on DrivingStereo [36]. The first example is taken around sunset with over-exposure. The second example is taken on a rainy day with under-exposure. In both challenge cases, SEDNet predicts accurate disparity.

of stereo matching, and a new loss to match the distribution of uncertainties with that of disparity errors. To generate the inputs to this new loss, we approximate the distributions from the samples of disparity errors and uncertainty values in a differentiable way via a soft-histogramming technique.

We present extensive experimental validation of SEDNet’s performance in disparity estimation and uncertainty prediction on large datasets with ground truth. SEDNet is superior to baselines with similar, even identical, architecture, but without the proposed loss function. Our main contributions are:

- a novel uncertainty estimation subnetwork that extracts information from the intermediate multi-resolution disparity maps generated by the disparity subnetwork,
- a differentiable soft-histogramming technique used to approximate the distributions of disparity errors and estimated uncertainties,
- a loss based on KL divergence applied on histograms obtained with the above technique.

2. Related Work

We refer readers to recent surveys on deep stereo matching [32] and on confidence estimation [30]. Here we summarize the most relevant publications to our work.

Stereo matching networks operate on a cost volume, which aggregates 2D features at each potential disparity for every pixel, and can be constructed via correlation or concatenation. Correlation-based networks such as DispNetC [25], iResNet [24] and SegStereo [38], generate a single-channel correlation map between features extracted from the two views at each disparity level, favoring computation efficiency at the expense of losing the structural and semantic information in the feature representation. Concatenation-based networks, such as GCNet [16], PSMNet [2] and GANet [39], assemble features from both views at the disparity specified by the corresponding element of the cost volume. This promotes learning of contextual features but requires more parameters and a subsequent aggregation network.

We select GwcNet [11] as the foundation of our network. GwcNet takes a hybrid approach by reducing the dimension of the unary feature channels before concatenation in the cost volume. This is accomplished by a *Group Wise Correlation layer*, which takes as input N_c unary feature channels, divides them into N_g groups, computes the correlation between channels in each group at all disparity levels, and uses the resulting correlation scores to form the cost volume. This reduces the size of the cost volume and the computational cost of 3D convolutions by a factor of $N_c : N_g$, with N_g much smaller than N_c , but still provides rich similarity-measure features to the disparity estimator.

Researchers have also focused on model reliability. In Bayesian Neural Networks (BNNs), different models are

sampled from the distribution of weights to estimate the mean and variance of the target distribution in an empirical manner, yielding estimates of uncertainty [7, 27]. Additional empirical strategies such as Bootstrapped Ensembles [23] and Monte Carlo Dropout [8] also sample from the distribution of weights. On the other hand, Graves [10] and Blundell et al. [1] proposed to replace the sampling with variational inference.

Due to the high cost of training BNNs, methods for modeling the uncertainty or confidence in a predictive manner have also attracted interest. We distinguish between confidence and uncertainty: confidence is a binary variable trained with the BCE loss, while uncertainty is a continuous variable trained with L1 or L2 loss. Nix and Weigend [28] introduce NNs with one output for model prediction and one for data noise (aleatoric uncertainty). In addition to aleatoric uncertainty which captures the data noise of the observations, epistemic uncertainty, which accounts for the uncertainty of the model parameters, can also be modeled [3]. To capture both types, Kendall et al. [14, 15] proposed to combine empirical and predictive methods in a joint framework.

CNNs have been used to estimate confidence in stereo matching. The Confidence CNN (CCNN) [31], Patch Based Confidence Prediction (PBCP) [33], the Early Fusion Network (EFN) and the Late Fusion Network (LFN) [4] and Multi Modal CNN (MMC) [5] only use small patches of the disparity maps. Conversely, the Global Confidence Network (ConfNet) and the Local-Global Confidence Network (LGC) [35] introduce U-Net like architectures and take both the image and disparity map as inputs. As a baseline, we use the Locally Adaptive Fusion Network (LAF) [18] which predicts the confidence map based on tri-modal inputs: the cost and disparity maps and the color image. An extension based on knowledge distillation has also been published [17]. These strategies are effective and cheaper than the empirical ones, since they only require one forward pass.

Only a subset of the confidence estimation literature has focused on joint disparity and confidence estimation. The Reflective Confidence Network (RCN) [34] is the first to combine a disparity and a confidence loss. The Unified Confidence Network (UCN) [19] and the Adversarial Confidence Network (ACN) [20] jointly estimate confidence and disparity from pre-computed cost volumes. UCN is self-supervised, while ACN combines a generative cost aggregation network and a discriminative confidence estimation network in an adversarial manner. Mehlretter [26] presents an approach that predicts both epistemic and aleatoric uncertainty using a Bayesian Neural Network, based on GCNet [16]. KL divergence is used to measure the distance between the approximation of the distribution of network parameters estimated by variational difference and the exact posterior distribution. (It should be noted that we use

KL divergence for a completely different purpose on the distribution of disparity errors.)

Relevant research in adjacent areas of computer vision includes the work of Poggi et al. [29] who comprehensively evaluate uncertainty estimation for self-supervised monocular depth estimation. Ilg et al. [13] study empirical ensembles, predictive models and predictive ensembles as uncertainty models for optical flow estimation.

3. Method

The objective of our work is to jointly estimate the disparity and its uncertainty. An important benefit of this joint formulation is that the multi-task network learns to predict more accurate disparities than the standalone disparity estimator when the uncertainty subnetwork is added. Given a stereo image pair $\mathbf{X} = \{\mathbf{x}_l, \mathbf{x}_r\}$, with image dimensions $H \times W$, and the corresponding ground truth disparity \mathbf{d} , the prediction $\hat{\mathbf{d}}$ of a stereo-matching network f_θ can be represented as $\hat{\mathbf{d}} = f_\theta(\mathbf{x}_l, \mathbf{x}_r)$. For each pixel i , the error $\epsilon^{(i)}$ of the prediction is calculated using the L1 loss.

Kendall and Gal [14] use the negative log-likelihood of the prediction model as the loss function to be minimized in pixel-wise tasks.

We take the formulation a step further by requiring that the network generate a distribution of uncertainties that matches the distribution of errors. To this end, we propose to minimize the divergence \mathcal{D} between the distributions of predicted uncertainty and actual disparity error.

In the following subsections, we present aleatoric uncertainty estimation (Section 3.1), the proposed KL divergence loss (Section 3.2), our network architecture (Section 3.3), and the combined loss function (Section 3.4).

3.1. Aleatoric Uncertainty Estimation

In order to predict uncertainty and reduce the impact of noise, Kendall and Gal [14] minimize the pixel-wise negative log-likelihood of the prediction model, assuming that it follows a Gaussian distribution. The subsequent work of Ilg et al. [13] shows that the predicted distribution can be modeled as either Laplacian or Gaussian depending on whether the L1 or L2 loss is used for disparity estimation. Since we use the former, we can write the prediction model as:

$$p(\mathbf{d}|f_\theta(\mathbf{X})) = \mathcal{Laplace}(f_\theta(\mathbf{X}), \sigma) \quad (1)$$

where the mean is given by the model output and σ is the observation noise scalar.

To model aleatoric uncertainty, Kendall and Gal [14] introduce pixel-specific noise parameters $\sigma^{(i)}$. We follow the approach of Ilg et al. [13], who do the same for a Laplacian model, and obtain the following pixel-wise loss function:

$$\mathcal{L}_{log} = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{\mathbf{d}}^{(i)} - \mathbf{d}^{(i)}|}{\exp(\mathbf{s}^{(i)})} + \frac{1}{n} \sum_{i=1}^n \mathbf{s}^{(i)} \quad (2)$$

where $\hat{\mathbf{d}}^{(i)}$ and $\mathbf{d}^{(i)}$ are the predicted and ground truth disparity for pixel i , $\mathbf{s}^{(i)}$ is the log of the observation noise scalar $\sigma^{(i)}$, and n is equal to the total number of the pixels. Equation (2) may be viewed as a robust loss function where the residual loss for a pixel is attenuated by its uncertainty, while the second term acts as a regularizer. We follow the authors' suggestion and train the network to predict the log of the observation noise scalar, s , for numerical stability.

3.2. Matching the Distribution of Errors

Training a model using Eq. (2) as the loss improves disparity estimation accuracy and favors uncertainty estimates correlated with the errors. Ideally, we would like each uncertainty estimate to be a precise predictor of the corresponding disparity error. Since this is infeasible, *we would like the distribution of uncertainties to match the distribution of errors.*

The Kullback-Leibler (KL) divergence [22] is a natural choice for measuring the dissimilarity between the distribution of ϵ and that of σ . Since the KL divergence is asymmetric, we choose the distribution of ϵ as the reference. Therefore, our network should also minimize the following objective function:

$$\mathcal{L}_{div} = \mathcal{D}(P_\sigma(d)||Q_\epsilon(d)) = \int_0^{d_{max}} P_\sigma(d) \log \frac{P_\sigma(d)}{Q_\epsilon(d)} dd \quad (3)$$

where d spans the disparity range. Since the network regresses disparity, the continuous formulation of KL divergence is appropriate.

Minimizing Eq. (3) directly requires closed form expressions for the two distributions, which are not available to us. They could be modeled as Laplace distributions, but the maximum likelihood estimator is not differentiable. Moreover, fitting models to the data may be imprecise at the tails of the distributions. Therefore we choose non-parametric representations in the form of histograms.

Histogramming is also not a differentiable operation, leading us to *soft-histogramming* as a differentiable alternative. We specify a set of bins for the histograms based on the statistics of the errors ϵ , since their distribution is the one that should be matched by the distribution of uncertainty estimates σ . Since the L1 loss, and our network in general, does not discriminate between positive and negative errors, we work with absolute values of ϵ and σ .

For each batch during training, we compute the mean and standard deviation of the error, μ_ϵ and b_ϵ , set $C_0 = \mu_\epsilon$ as the center of the first bin, and $C_m = \mu_\epsilon + \alpha_m b_\epsilon$ as the center of the last bin. (We use b_ϵ for the standard deviation to avoid overloading σ or s . Also note that the last bin extends to the disparity range, which is also the maximum possible error.) We then define $m - 1$ centers evenly spaced in a linear or logarithmic scale between the first and last center.

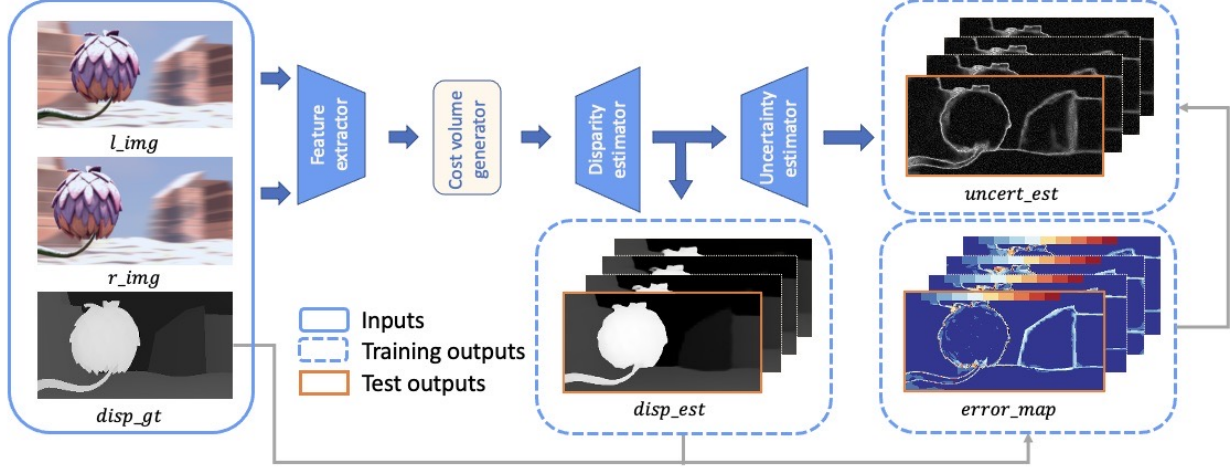


Figure 2. An illustration of SEDNet. The stereo matching subnetwork consists of the feature extractor, cost volume generator and disparity estimator. It takes a rectified pair of images as input and predicts disparity maps at multiple resolutions. The uncertainty estimation subnetwork takes the predicted disparity maps as input and predicts corresponding uncertainty maps. The error maps, between the ground truth and predicted disparity, are used to supervise uncertainty estimation. During training, the network keeps the output at all resolutions, but returns only the highest resolution disparity and uncertainty maps during testing.

Given the bin centers, we compute a soft-histogram for the errors and one for the uncertainties as follows, considering all pixels with ground truth disparity, and error values. We present the steps for ϵ here. For each error $\epsilon^{(i)}$, we compute weights for every bin center which are inversely proportional to the distance.

$$w_j(\epsilon^{(i)}) = \lambda_1 \cdot \exp\left(-\frac{(\mu_\epsilon + \alpha_j b_\epsilon - \epsilon^{(i)})^2}{\lambda_2}\right) \quad (4)$$

where λ_1 and λ_2 are hyper-parameters. Softmax is then applied to favor the nearest bins and the contributions of n pixels are accumulated in the bins of the histogram H_ϵ .

$$H_\epsilon(j) = \frac{1}{n} \sum_{i=1}^n \frac{\exp(w_j(\epsilon^{(i)}))}{\sum_{j=0}^m \exp(w_j(\epsilon^{(i)}))}, \quad j \in [0, m] \quad (5)$$

The histogram for σ , H_σ , is obtained similarly. (Note that the bins of both histograms are defined in terms of b_ϵ .)

The loss representing the discrete form of the KL divergence between the two histograms is given by:

$$\mathcal{L}_{div} = \sum_{j=0}^m H_\epsilon(j) \log \frac{H_\epsilon(j)}{H_\sigma(j)} \quad (6)$$

3.3. SEDNet

Our network architecture, named **SEDNet**, includes a disparity estimation subnetwork, an uncertainty estimation subnetwork, and is shown in Figure 2. In all experiments, we have adopted GwcNet [11] as the disparity estimation subnetwork, among other options, and we have designed

a novel uncertainty estimation subnetwork that interfaces with GwcNet. It is worth noting that the uncertainty subnetwork is *extremely small*.

The GwcNet subnetwork extracts features from the images using a ResNet-like feature extractor, generates the cost volume, and assigns disparities to pixels using the soft-argmax operator [16]. The output module of the disparity predictor generates K disparity maps at different resolutions.

We propose a new uncertainty estimator integrated with the stereo matching network. The uncertainty estimation subnetwork learns to predict the log of the observation noise scalar, the error, at each pixel. The proposed subnetwork takes the multi-resolution disparity predictions as input, computes the *pairwise differences vector (PDV)* and passes it to a pixel-wise MLP to regress the uncertainty maps. Specifically, the disparity estimator outputs K disparity maps at different resolutions $\hat{\mathbf{d}} = \{\hat{\mathbf{d}}_1, \dots, \hat{\mathbf{d}}_K\}$, which are first upsampled to full-resolution and then undergo pairwise differencing to form the PVD, which consists of $\binom{K}{2}$ elements. The output set of uncertainty maps \mathbf{S} also contains K resolutions to match the disparity maps. We use $K = 4$ in all experiments.

3.4. Loss Function

Our loss function combines two parts: (1) the log-likelihood loss to optimize the error and uncertainty, (2) the KL divergence loss to match the distribution of uncertainty with the error. The total loss considers all disparity and un-

certainty maps upsampled to the highest resolution:

$$\mathcal{L} = \sum_{k=1}^K c_k \cdot (\mathcal{L}_{log,k} + \mathcal{L}_{div,k}) \quad (7)$$

where c_k denotes the coefficients for the k^{th} resolution level, $\mathcal{L}_{log,k}$ and $\mathcal{L}_{div,k}$ are computed by Eq. (2) and Eq. (6) on the prediction of the corresponding resolution level.

4. Experiment Results

In this section, we present our experimental setup and results on *within-domain* and *cross-domain* experiments. Datasets, evaluation metrics and baselines are described in Section 4.1, implementation details in Section 4.2, and experimental results in Sections 4.3 and 4.4. Additionally, a synthetic-to-real transfer evaluation is presented in Section 4.5. We provide more quantitative and qualitative results, extending Section 4, in the Supplement, which also includes ablation studies on several aspects of SEDNet and the baselines, as well as disparity, error, and uncertainty maps of difficult examples.

4.1. Datasets and evaluation metrics

SceneFlow [25] is a collection of three synthetic stereo datasets: FlyingThings3D, Driving, Monkaa. The datasets provide 35,454 training and 4,370 test stereo pairs in 960×540 pixel resolution with dense ground-truth disparity maps. We use the *finalpass* versions of the rendered images which are more realistic because of the motion blur and depth of field effect.

Virtual KITTI 2 (VK2) [6] is a synthetic clone of KITTI [9]. It consists of 21,260 synthetic $1,242 \times 375$ stereo pairs from 6 driving scenes with 10 different imaging and weather conditions. Scene 006 (VK2-S6) is specified as the test set by the authors of the dataset. Since the car with the cameras is stopped for a long part of Scene 006 and only other cars move in the images, we split the last part of the scene where the car moves and denote it as *VK2-S6-Moving*. We report results separately on this subset. Since results on VK2-S6-Moving do not suffer from bias due to the almost constant background of the first part of the scene, evaluation on VK2-S6-Moving is more informative and fair.

DrivingStereo [37] is a large real-world autonomous driving dataset. It contains 174,437 training and 7,751 test stereo pairs at 881×400 pixel resolution. The dataset provides sparse ground truth disparity as well as a challenging subset (*DS-Weather*) of 2,000 stereo pairs in 4 different weather conditions.

For all datasets above, we exclude pixels with disparities $d > 192$ in training.

Metrics. To evaluate disparity estimation, we compute the endpoint error (**EPE**) and the percentage of outliers (**D1**)

(i.e., the percentage of pixels with $EPE > 3px$ or $\geq 5\%$ of the true depth).

To evaluate uncertainty estimation, we use density-EPE ROC curves and the area under the curve (**AUC**) [12,18,30]. The ROC curves in our case measure EPE (not binary correctness) at increasing disparity map density by successively adding pixels in increasing order of uncertainty. The optimal AUC is obtained by adding pixels in order of increasing EPE and is therefore the lowest value any algorithm could achieve given the set of EPE values, while the estimated AUC is sorted by predicted uncertainty. To evaluate the precision of uncertainty estimation, we also introduce the absolute prediction error (**APE**), which is the average L1 distance between the error and the observation noise scalar, σ . We report the average and median APE over all pixels.

Baselines. Since we use GwcNet [11] as the disparity estimation subnetwork of SEDNet, we compare SEDNet with three baselines: (1) the original GwcNet trained with smooth L1 loss. (2) LAF-Net [18] trained under the BCE loss on the left RGB images, the cost volumes and predicted disparity maps of GwcNet at the highest resolution. (3) SEDNet but only trained with the log-likelihood loss, therefore similar to Kendall and Gal’s model [14]. We use \mathcal{L}_{log} in tables and figures for this baseline. We selected GwcNet because it is the backbone of SEDNet, and LAF-Net as the confidence estimation baseline due to its strong performance within the training domain according to [30].

4.2. Implementation Details

We implemented all networks in PyTorch and used the Adam optimizer [21] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for all experiments. Training of all models was stopped before overfitting occurred.

Experiments on the VK2 dataset were performed on two NVIDIA RTX A6000 GPUs, each with 48 GB of RAM. For this dataset, we trained all models from scratch with an initial learning rate of 0.0001, down-scaled by 5 every 10 epochs. During training, we randomly cropped 512×256 patches from the images. During testing, we evaluated at the full resolution of VK2.

Experiments on the DrivingStereo dataset (DS) were also performed on two NVIDIA RTX A6000 GPUs. For this dataset, we did two experiments using the models pre-trained on VK2: (1) we finetuned on the DS training set with a learning rate starting from 0.0001, down-scaled by 2 every 3 epochs after epoch 10, then performed in-domain evaluation on the DS test set; (2) we skipped the finetuning step and performed cross-domain evaluation on DS-Weather subset. During training, we randomly cropped the inputs to be the same size as in the VK2 experiments. During testing, we padded the test samples to be the same resolution as VK2.

Dataset	Method	Loss						Inliers		Disparity \downarrow		APE \downarrow		AUC \downarrow	
		BCE	L1	Log	KL	Bins	Scale	Def.	Pct(%)	EPE	D1(%)	Avg.	Median	Opt.	Est.
Scene Flow	GwcNet	-	✓	-	-	-	-	-	-	0.7758	4.127	-	-	10.9291	-
	+LAF	✓	-	-	-	-	-	-	-	0.7758	4.127	-	-	10.9291	20.0813
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE <5	96.96	0.7611	4.131	0.6999	0.0728	5.7449	12.1121
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+3b_\epsilon$	98.42	0.6754	3.963	0.5797	0.0432	4.9134	8.7195
VK2-S6	GwcNet	-	✓	-	-	-	-	-	-	0.4125	1.763	-	-	6.0962	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE <5	98.86	0.3899	1.584	0.4136	0.1753	4.6872	12.5320
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+3b_\epsilon$	99.24	0.3109	1.392	0.5234	0.1454	4.1726	9.7637
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+5b_\epsilon$	99.68	0.3236	1.427	0.3561	0.1096	4.2767	9.9843
VK2-S6-Moving	GwcNet	-	✓	-	-	-	-	-	-	0.4253	1.689	-	-	5.9184	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE <5	98.91	0.4231	1.537	0.4575	0.1890	4.3663	11.3532
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+3b_\epsilon$	99.62	0.3577	1.389	0.5958	0.1573	3.9012	8.8339
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+5b_\epsilon$	99.76	0.3862	1.420	0.4002	0.1164	4.0423	9.0631
DrivingStereo	+ \mathcal{L}_{log} (FT)	-	✓	✓	-	-	-	-	-	0.5332	0.2641	0.3449	0.2297	21.7002	45.7096
	+SEDNet(FT)	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+5b_\epsilon$	99.86	0.5264	0.2439	0.3324	0.2267	21.2856	44.3297
DS-Weather	GwcNet	-	✓	-	-	-	-	-	-	1.6962	8.313	-	-	44.4896	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE <5	95.78	2.3944	6.666	2.1443	0.4383	41.1909	95.4264
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+3b_\epsilon$	98.95	1.5637	6.508	2.3406	0.5309	38.4871	86.1118
	+SEDNet	-	✓	✓	✓	11	log	EPE $<\mu_\epsilon+5b_\epsilon$	99.41	1.7051	6.057	1.5842	0.6104	39.8057	87.1882

Table 1. Quantitative results: (1) *within-domain* on SceneFlow, VK2-S6 and VK2-S6-Moving; (2) after finetuning (FT) on DrivingStereo; (3) *cross-domain* on DS-Weather. The best results in each category in each experiment are in bold typeface. SEDNet outperforms the baselines with respect to disparity and uncertainty metrics in the majority of experiments.

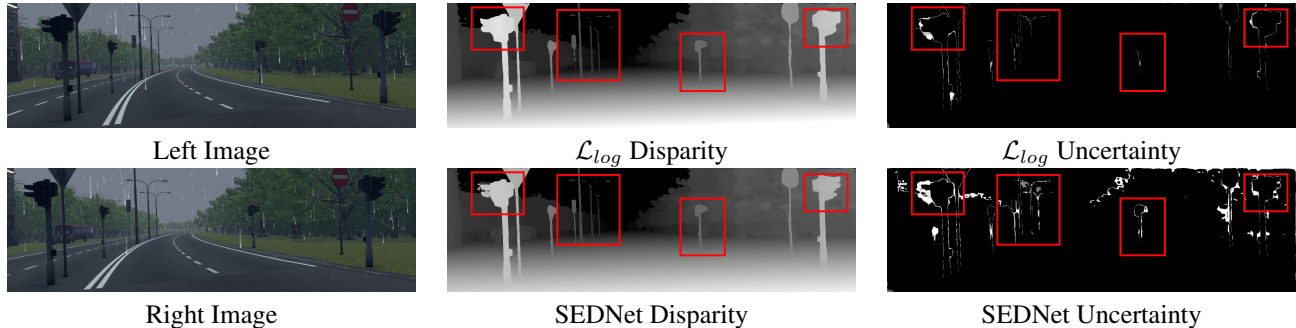


Figure 3. Example from VK2-S6. In a rainy scene with poor illumination, recognizing objects far away from the camera is even difficult for human observers. \mathcal{L}_{log} fails to predict the disparity of objects, such as the traffic sign near the left street light. On the other hand, SEDNet accurately predicts the disparity of these challenging objects, while its uncertainty map also captures more information, such as the trees in the background, the traffic light near the trees and the street lights.

Experiments on the SceneFlow dataset were performed on an Nvidia TITAN RTX GPU with 24 GB memory. We trained all models from scratch on 256×128 patches cropped from half-resolution images to limit memory consumption. We set the initial learning rate to 0.001 and down-scaled it by 2 every 2 epochs after epoch 10.

For all the experiments above, we applied an inlier filtering strategy during training, which only back-propagates from the inliers. See the last paragraph of Section 4.3 for

Architecture	Params	MACs(G)
GwcNet	6,909,728	1075.82
SEDNet	6,909,918	1075.91

Table 2. Comparison on number of parameters and computational complexity. MAC stands for multiply-accumulate operations. SEDNet only adds a 3-layer MLP, with 190 parameters, as an uncertainty decoder to GwcNet.

details. Table 2 shows the number of parameters and computational complexity of GwcNet and SEDNet.

4.3. Qualitative and Quantitative Results

In Table 1, we present results: (1) *within-domain* on SceneFlow and VK2; (2) on the DrivingStereo test set after finetuning the VK2 models on the DrivingStereo training set; (3) *cross-domain* on the DS-Weather challenge test set by directly applying the model trained on VK2 without finetuning.

Disparity Estimation. In all experiments, SEDNet achieves lower errors than all the baselines. See the EPE and D1 columns in Table 1. Even in extreme weather like fog and rain, SEDNet predicts good disparity unaffected by poor illumination and blur. See Figure 3.

Uncertainty Estimation. Our method outperforms the baselines in all experiments according to the AUC metric,

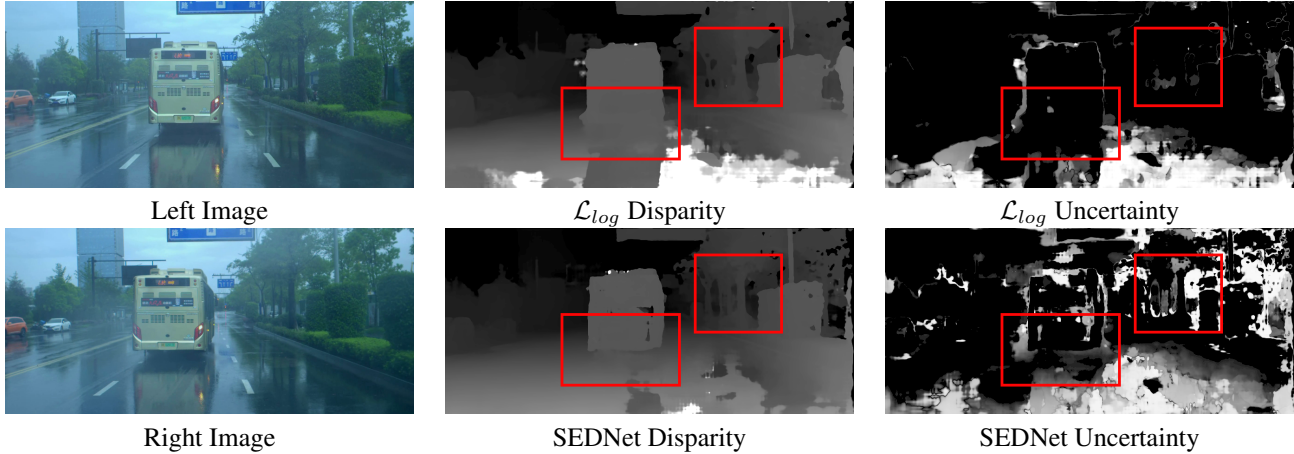


Figure 4. Example from DS-Weather. Unlike the synthetic data, the rainy-day real images do not only suffer from poor illumination, but also pose challenges due to reflections in the water. In this example, the road is like a mirror, misleading the \mathcal{L}_{log} model. Recall that the LIDAR ground truth disparity is very sparse, and is even sparser in reflective regions. The disparity map of \mathcal{L}_{log} fails to distinguish the car and the reflection, but SEDNet is able to estimate the correct disparity and the uncertainty of the car.

as shown in the last two columns in Table 1. Compared to \mathcal{L}_{log} , SEDNet decreases the estimated AUC by 20% – 30% in the in-domain experiments, with a 10% decrease in optimal AUC, which depends on EPE. In the cross-domain evaluation, the advantage of SEDNet is even more evident. Figure 4 shows uncertainty maps for real data, on which our method captures details more faithfully. The ROC curves of the best \mathcal{L}_{log} and SEDNet models based on EPE on VK2-S6-Moving are presented in Figure 5.

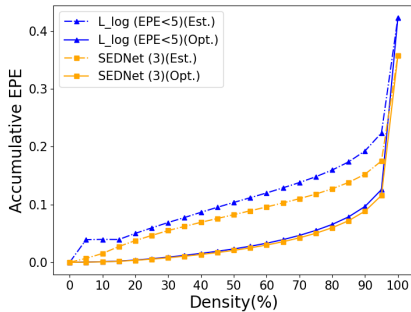


Figure 5. Comparison of density-EPE curves on VK2-S6-Moving. "L_log (EPE<5)" and "SEDNet (3)" represent the \mathcal{L}_{log} (with fixed inliers) and the SEDNet (with adaptive inliers of $EPE < \mu_e + 3b_e$) of the VK2-S6-Moving experiment in Table 1. While their corresponding optimal AUCs are almost equal, the advantage of SEDNet in estimated AUC is significant.

Effects of Back-propagation from Inliers. We apply two kinds of inlier filters: one with a fixed threshold that excludes all pixels that have an EPE larger than 5 from back-propagation; and one with adaptive threshold which excludes pixels that have EPE greater than a specified number of b_e from the mean error. Back-propagation from the

inliers only helps the network improve its performance on both disparity and uncertainty estimation. We attribute this to the suppression of harmful outliers that give rise to large gradients. Quantitative results for the baselines and the proposed method with different inlier settings are reported in the Supplement. The results show that using adaptive thresholds is better than fixed thresholds. Fixed thresholds exclude more pixels, especially at lower resolutions and in the early stages of training, preventing the network from learning how to correct them.

4.4. Matching the Error Distribution.

Comparing the results of APE in all experiments in Table 1 and Figure 6, the difference between the errors and the predicted uncertainty is always close to the EPE, as expected. SEDNet, however, achieves lower APE, showing

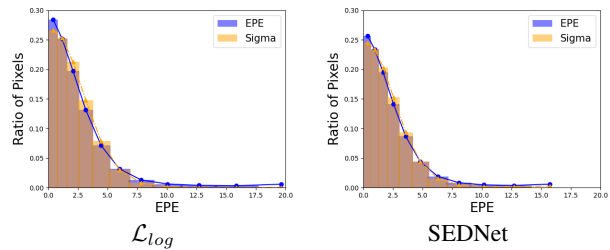


Figure 6. Distribution of error and predicted uncertainty. We pick the best model of \mathcal{L}_{log} and SEDNet in VK2-S6-Moving, then randomly sample 5,000,000 points from the outputs of the two models. The distributions of error and σ of SEDNet are much closer than those of the \mathcal{L}_{log} model, especially in the first 5 bins that contain more than 95% of the samples. The shorter tail of SEDNet also indicates fewer gross errors.

Dataset	Method	Loss						Inliers		Disparity↓		APE↓		AUC↓	
		BCE	L1	Log	KL	Bins	Scale	Def.	Pct(%)	EPE	D1(%)	Avg.	Median	Opt.	Est.
VK2-S6-Morning	GwcNet	-	✓	-	-	-	-	-	-	0.4642	1.740	-	-	6.1845	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	98.82	0.4774	1.624	0.5067	0.1872	4.6698	12.5192
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.62	0.4003	1.442	0.6183	0.1553	4.1847	9.4063
VK2-S6-Sunset	GwcNet	-	✓	-	-	-	-	-	-	0.4810	1.825	-	-	6.6907	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	98.84	0.4863	1.627	0.5060	0.1827	5.0075	13.7848
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.61	0.4108	1.475	0.6189	0.1509	4.5840	10.7946
VK2-S6-Fog	GwcNet	-	✓	-	-	-	-	-	-	0.4660	1.812	-	-	6.8355	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	98.98	0.4425	1.448	0.4609	0.1865	4.8983	12.1305
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.71	0.3731	1.288	0.5517	0.1547	4.4200	9.9380
VK2-S6-Rain	GwcNet	-	✓	-	-	-	-	-	-	0.4618	1.700	-	-	6.6774	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	98.88	0.4707	1.571	0.4899	0.1861	4.9351	13.3214
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.69	0.3873	1.356	0.6685	0.1537	4.4013	10.3362
DS-Cloudy	GwcNet	-	✓	-	-	-	-	-	-	1.3413	5.229	-	-	37.4263	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	97.48	1.4780	3.948	1.2617	0.3513	34.4488	82.5380
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	98.83	1.3183	4.414	1.5260	0.4021	33.9037	73.6330
DS-Sunny	GwcNet	-	✓	-	-	-	-	-	-	1.5448	6.991	-	-	38.7386	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	97.08	1.4837	4.631	1.2806	0.3835	35.5226	85.8715
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	98.64	1.5548	5.878	3.0025	0.4808	35.6523	83.2573
DS-Foggy	GwcNet	-	✓	-	-	-	-	-	-	1.5476	8.859	-	-	51.4640	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	94.89	2.9553	9.015	2.6923	0.5556	48.7136	101.7025
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.27	1.5398	7.357	2.4109	0.7023	47.7932	97.8627
DS-Rainy	GwcNet	-	✓	-	-	-	-	-	-	3.1918	17.356	-	-	68.0346	-
	+ \mathcal{L}_{log}	-	✓	✓	-	-	-	EPE<5	98.79	5.3539	12.501	4.9480	0.5759	59.3952	146.8906
	+SEDNet	-	✓	✓	✓	11	log	EPE< $\mu_\epsilon+3b_\epsilon$	99.10	2.2165	11.020	2.6599	0.6722	50.8103	110.8360

Table 3. Quantitative results of synthetic to real evaluation. The top 4 subsets from VK2-S6-Moving are synthetic datasets, while the rest 4 subsets from DS-Weather are the real dataset. The best results in each category in each experiment are in bold typeface. SEDNet outperforms the baselines especially on the uncertainty estimation of the real data and under terrible weather (i.e., foggy and rainy).

that it matches the true distribution better.

4.5. Generalization from Synthetic to Real Data

Stereo-matching networks are typically trained on synthetic data and fine-tuned on small amounts of data from the target domain due to the cost and difficulty of acquiring real data with ground truth depth. In this section, we extend the experiments of VK2-S6 and DS-Weather in Table 1 to compare the generalization performance on unseen real domains of all methods trained only on synthetic data. We picked four synthetic subsets from VK2-S6, specifically *Morning*, *Sunset*, *Fog* and *Rain*, that have similar illumination conditions, visibility level and weather with the four real subsets of DS-Weather [37], i.e., *Cloudy*, *Sunny*, *Foggy* and *Rainy*. VK2-S6-Morning and VK2-S6-Sunset have similar illumination to DS-Cloudy and DS-Sunny, but the latter two are more challenging due to camera underexposure and overexposure. Even though they are acquired under the same weather, DS-Foggy and DS-Rainy are more difficult than VK2-S6-Fog and VK2-S6-Rain. This can be seen by comparing Figures 3 and 4. The synthetic examples only mimic the poor lighting conditions and challenges caused by the fog and rain, but ignore the Tyndall effect and reflections caused by the fog and stagnant water.

As mentioned above, all models are trained only on Scenes 001, 002, 018 and 020 of VK2. Quantitative results on synthetic to real transfer are reported in Table 3, where we only report the best model of each method. The top-

performing variant of SEDNet outperforms the baselines in the majority of experiments. An extended version of this table can be found in Table S.2 in the Supplement.

5. Conclusion

We have presented a novel approach for joint disparity and uncertainty estimation from stereo image pairs. The key idea is a unique loss function based on the KL divergence between the distributions of disparity errors and uncertainty estimates. This is made possible by a differentiable histogramming scheme that we also introduce here. To implement our approach, we extended the GwcNet architecture to include an uncertainty estimation subnetwork with only 190 parameters. Our experiments on multiple large datasets have demonstrated that our approach, named SEDNet, is effective in both disparity and uncertainty prediction. The success of our method is attributed to the novel loss function. SEDNet easily surpasses GwcNet in disparity estimation even though they have essentially the same capacity and almost identical architecture, up to the tiny uncertainty estimation subnetwork. We are optimistic that our approach will be similarly successful in other pixel-wise regression tasks, which we plan to address in future research.

Acknowledgment. This research has been supported in part by the National Science Foundation under award 2024653.

References

- [1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, pages 1613–1622, 2015. 2
- [2] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *CVPR*, pages 5410–5418, 2018. 2
- [3] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2
- [4] Zehua Fu, Mohsen Ardabilian, and Guillaume Stern. Stereo matching confidence learning based on multi-modal convolution neural networks. In *International Workshop on Representations, Analysis and Recognition of Shape and Motion From Imaging Data*, pages 69–81. Springer, 2017. 2
- [5] Zehua Fu and Mohsen Ardabilian Fard. Learning confidence measures by multi-modal convolutional neural networks. In *WACV*, pages 1321–1330, 2018. 2
- [6] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, pages 4340–4349, 2016. 5
- [7] Yarin Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016. 2
- [8] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, pages 1050–1059, 2016. 2
- [9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012. 5
- [10] Alex Graves. Practical variational inference for neural networks. In *NeurIPS*, volume 24, 2011. 2
- [11] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *CVPR*, 2019. 2, 4, 5
- [12] Xiaoyan Hu and Philippos Mordohai. A quantitative evaluation of confidence measures for stereo vision. *PAMI*, 34(11):2121–2133, 2012. 1, 5
- [13] Eddy Ilg, Ozgun Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *ECCV*, pages 652–667, 2018. 3
- [14] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584, 2017. 1, 2, 3, 5
- [15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 2
- [16] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *ICCV*, pages 66–75, 2017. 2, 4
- [17] Sunok Kim, Seungryong Kim, Dongbo Min, Pascal Frossard, and Kwanghoon Sohn. Stereo confidence estimation via locally adaptive fusion and knowledge distillation. *PAMI*, 2022. 2
- [18] Sunok Kim, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. LAF-Net: Locally adaptive fusion networks for stereo confidence estimation. In *CVPR*, 2019. 2, 5
- [19] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Unified confidence estimation networks for robust stereo matching. *TIP*, 28(3):1299–1313, 2018. 1, 2
- [20] Sunok Kim, Dongbo Min, Seungryong Kim, and Kwanghoon Sohn. Adversarial confidence estimation networks for robust stereo matching. *TIST*, 22(11):6875–6889, 2020. 1, 2
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5
- [22] S Kullback and R Leibler. On information and sufficiency. *Annals of mathematical statistics*, 22, 79–86. *MathSciNet MATH*, 1951. 3
- [23] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, volume 30, 2017. 2
- [24] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *CVPR*, pages 2811–2820, 2018. 2
- [25] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, pages 4040–4048, 2016. 2, 5
- [26] Max Mehlretter. Joint estimation of depth and its uncertainty from stereo images using bayesian deep learning. *ISPRS*, 2:69–78, 2022. 1, 2
- [27] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. 2
- [28] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *ICNN*, volume 1, pages 55–60. IEEE, 1994. 2
- [29] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020. 3
- [30] Matteo Poggi, Seungryong Kim, Fabio Tosi, Sunok Kim, Filippo Aleotti, Dongbo Min, Kwanghoon Sohn, and Stefano Mattoccia. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation. *PAMI*, 2021. 1, 2, 5
- [31] Matteo Poggi and Stefano Mattoccia. Learning from scratch a confidence measure. In *BMVC*, 2016. 2
- [32] Matteo Poggi, Fabio Tosi, Konstantinos Batsos, Philippos Mordohai, and Stefano Mattoccia. On the synergies between machine learning and binocular stereo for depth estimation from images: a survey. *PAMI*, 44(9):5314–5334, 2021. 2
- [33] Akihito Seki and Marc Pollefeys. Patch based confidence prediction for dense disparity map. In *BMVC*, pages 23.1–23.13, 2016. 2
- [34] Amit Shaked and Lior Wolf. Improved stereo matching with constant highway networks and reflective confidence learning. In *CVPR*, pages 4641–4650, 2017. 1, 2

- [35] Fabio Tosi, Matteo Poggi, Antonio Benincasa, and Stefano Mattoccia. Beyond local reasoning for stereo confidence estimation with deep learning. In *ECCV*, pages 319–334, 2018. [2](#)
- [36] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019. [1](#)
- [37] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *CVPR*, 2019. [5](#), [8](#)
- [38] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. SegStereo: Exploiting semantic information for disparity estimation. In *ECCV*, pages 636–651, 2018. [2](#)
- [39] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In *CVPR*, 2019. [2](#)