Parameterized Cost Volume for Stereo Matching Supplementary Materials

Jiaxi Zeng[1]†, Chengtang Yao[1,3]†, Lidong Yu[3], Yuwei Wu[1]*, Yunde Jia[2],
[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
[3]Autonomous Driving Algorithm, Deeproute
{jiaxi,yao.c.t,wuyuwei,jiayunde}@bit.edu.cn
yvlidong@gmail.com

1. Derivation of Formulas

1.1. Derivation of Lemma 1

Lemma 1 *Given two Gaussian distribution $\mathcal{N}_p$ and $\mathcal{N}_q$, the KL divergence $F(\mathcal{N}_p||\mathcal{N}_q)$ is*

$$F(\mathcal{N}_p||\mathcal{N}_q) = log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}. \tag{1}$$

Proof. Assuming that $p(x) \sim \mathcal{N}_p(\mu_p, \sigma_p^2)$ and $q(x) \sim \mathcal{N}_q(\mu_q, \sigma_q^2)$, the KL divergence $F(\mathcal{N}_p||\mathcal{N}_q)$ can be formulated as

$$\begin{aligned} &F(\mathcal{N}_p||\mathcal{N}_q) \\ &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log q(x) dx. \end{aligned} \tag{2}$$

The first term of Eq 2 can be simplified as

$$\begin{aligned} &\int p(x) \log p(x) dx \\ &= \int p(x) \log(\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp(-\frac{(x-\mu_p)^2}{2\sigma_p^2}))dx \\ &= -\frac{1}{2}\log(2\pi\sigma_p^2)\int p(x)dx - \frac{1}{2\sigma_p^2}\int (x-\mu_p)^2 p(x)dx \\ &= -\frac{1}{2}\log(2\pi\sigma_p^2) - \frac{1}{2}. \end{aligned} \tag{3}$$

The second term of Eq 2 can be simplified as

$$\begin{aligned} &\int p(x) \log q(x) dx \\ &= \int p(x) \log(\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp(-\frac{(x-\mu_q)^2}{2\sigma_q^2}))dx \\ &= -\frac{1}{2}\log(2\pi\sigma_q^2)\int p(x)dx - \frac{1}{2\sigma_q^2}\int (x-\mu_q)^2 p(x)dx \\ &= -\frac{1}{2}\log(2\pi\sigma_q^2) - \frac{1}{2\sigma_q^2}\int [(x-\mu_p)^2 + \\ &\quad 2(\mu_p-\mu_q)(x-\mu_p) + (\mu_p-\mu_q)^2]p(x)dx \\ &= -\frac{1}{2}\log(2\pi\sigma_q^2) - \frac{\sigma_p^2 + (\mu_p-\mu_q)^2}{2\sigma_q^2}. \end{aligned} \tag{4}$$

Substituting the Eq 3 and Eq 4 into Eq 2, the KL divergence $F(\mathcal{N}_p||\mathcal{N}_q)$ is obtained as

$$F(\mathcal{N}_p||\mathcal{N}_q) = log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}. \tag{5}$$

1.2. Derivation of Lemma 2

Lemma 2 *Given two multi-Gaussian distribution $P = \sum_{i=1}^{i=M} \alpha_i^p \mathcal{N}_i^p$ and $Q = \sum_{i=1}^{i=M} \alpha_i^q \mathcal{N}_i^q$, the compact upper bound of KL divergence $F(P||Q)$ is*

$$F(P||Q) \leq \sum_{i=1}^{i=M} F(\alpha_i^p||\alpha_i^q) + \sum_{i=1}^{i=M} \alpha_i^p F(\mathcal{N}_i^p||\mathcal{N}_i^q). \tag{6}$$

Proof. According to the log sum inequality [1], for non-negative numbers $a_1, a_2, ..., a_m$ and $b_1, b_2, .., b_m$:

$$(\sum_{i=1}^{n} a_i) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \leq \sum_{i=1}^{n} a_i \log \frac{a_i}{b_i}. \tag{7}$$

---

† These authors contributed equally to this work.
* Corresponding author.

with equality if and only if $\frac{a_i}{b_i}$ is a constant. The inequation 6 can be derivated as follows:

$$
\begin{aligned}
F(P||Q) &= \int \sum_{i=1}^{i=M} \alpha_i^p \mathcal{N}_i^p \log \frac{\sum_{i=1}^{i=M} \alpha_i^p \mathcal{N}_i^p}{\sum_{i=1}^{i=M} \alpha_i^q \mathcal{N}_i^q} \\
&\leq \int \sum_{i=1}^{M} \alpha_i^p \mathcal{N}_i^p \log \frac{\alpha_i^p \mathcal{N}_i^p}{\alpha_i^q \mathcal{N}_i^q} \\
&= \sum_{i=1}^{M} \alpha_i^p \log \frac{\alpha_i^p}{\alpha_i^q} \int \mathcal{N}_i^p + \sum_{i=1}^{M} \alpha_i^p \int \mathcal{N}_i^p \log \frac{\mathcal{N}_i^p}{\mathcal{N}_i^q} \\
&= \sum_{i=1}^{M} \alpha_i^p \log \frac{\alpha_i^p}{\alpha_i^q} + \sum_{i=1}^{M} \alpha_i^p \int \mathcal{N}_i^p \log \frac{\mathcal{N}_i^p}{\mathcal{N}_i^q} \\
&= \sum_{i=1}^{i=M} F(\alpha_i^p || \alpha_i^q) + \sum_{i=1}^{i=M} \alpha_i^p F(\mathcal{N}_i^p || \mathcal{N}_i^q).
\end{aligned} \tag{8}
$$

2. Architecture of Uncertainty-aware Refinement Module

As shown in Figure 1, we use two convolutional layers with leaky-ReLU to regress the uncertainty map U with the weighted variances $\alpha^t \times \sigma^T$ and disparity map $\bar{\mu}$ as the inputs. Subsequently, the uncertainty map is concatenated with the disparity map and the left features to feed into a 4-layer dilated convolutional network, regressing the residual map R. Finally, we execute a disparity fusion process to obtain the final disparity map according to the Eq 11 of the main text.

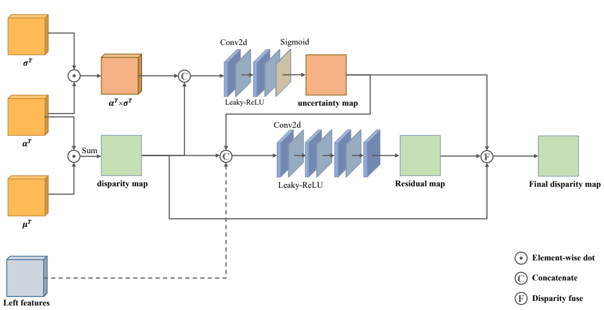

Figure 1. The structure of uncertainty-aware refinement module.

3. Ablation Study Details

As mentioned in Section 4.3 of the main text, we implement three different methods to compare with our parameterized cost volume. The first is the single Gaussian distribution with fixed variance (SGFV). We take the RAFT-stereo [4] as the implementation of it.

The second one is the single Gaussian distribution with an adaptive variance (SGAV). The initial disparity μ^0 is set to 0, and the initial variance σ^0 is 8. For the t-th iteration, the disparity candidates $D^t = \{d_1^t, d_2^t, ..., d_N^t\}$ are sampled from a range $[\mu^t - k\sigma^t, \mu^t + k\sigma^t]$ uniformly where k is 0.5 and N is 9. We calculate the costs based on the disparity candidates and then use the costs to predict the offsets $O^t = [o_1^t, o_2^t, ..., o_N^t]$ and probabilities $P^t = [p_1^t, p_2^t, ..., p_N^t]$ for the candidates. The thin volume is constructed as $V = [(d_1^t + o_1^t, p_1^t), (d_2^t + o_2^t, p_2^t), ..., (d_N^t + o_N^t, p_N^t)]$. We regress the disparity d^{t+1} and σ^{t+1} as follows:

$$
\mu^{t+1} = \sum_{i=1}^{N} p_i^t (d_i^t + o_i^t),
$$

$$
\sigma^{t+1} = \sum_{i=1}^{N} p_i^t (d_i^t + o_i^t - \mu^{t+1})^2.
$$

σ^{t+1} is clipped to avoid numerical explosion. d^{t+1} and σ^{t+1} are used for the next iteration. The μ^{t+1} and the $d_i^t + o_i^t$ are supervised to close to the ground truth.

The third method is based on the multiple Gaussian distribution with fixed variance (MGFV). We initialize 4 points $d \in \{0, 64, 128, 192\}$ to search for the ground truth. For each point, we use the sampling strategy of RAFT-Stereo to sample the disparity candidates for each Gaussian distribution and predict the updates of means and weights. The weighted average of the means at the last iteration is regarded as the final disparity. All the methods use the L1 loss used in RAFT-Stereo [4] to supervise the disparity sequences. We set the number of iterations to 6 for training and 4 for testing.

4. Supplementary Results of Experiments

4.1. Uncertainty-aware Refinement

We visualize the error maps of the last iteration and the uncertainty maps in Figure 2. It can be seen from the white boxes that regions with high uncertainty are generally correlated with large errors, which demonstrates the uncertainty-aware characteristics of our refinement module.

Method	D1-bg (%)	D1-fg (%)	D1-all (%)	Time (ms)
PCW-Net [9]	1.29	2.93	1.53	440
DeepPruner(best) [2]	1.71	3.18	1.95	180
CREStereo [3]	1.33	2.60	1.54	410
RAFT-Stereo [4]	1.45	2.94	1.69	380
AANet+ [11]	1.49	3.66	1.85	60
DeepPruner(fast) [2]	2.13	3.43	2.35	60
HITNet [10]	1.54	2.72	1.74	20
Dec-Net [13]	1.89	3.53	2.16	50
PCVNet (ours)	1.56	2.98	1.8	56

Table 1. The comparison of algorithms on non-occluded pixel areas in the KITTI 2015 dataset [5].

4.2. KITTI2015

In Table 1, we present the results of different methods on non-occluded pixel areas. As shown in the table, our

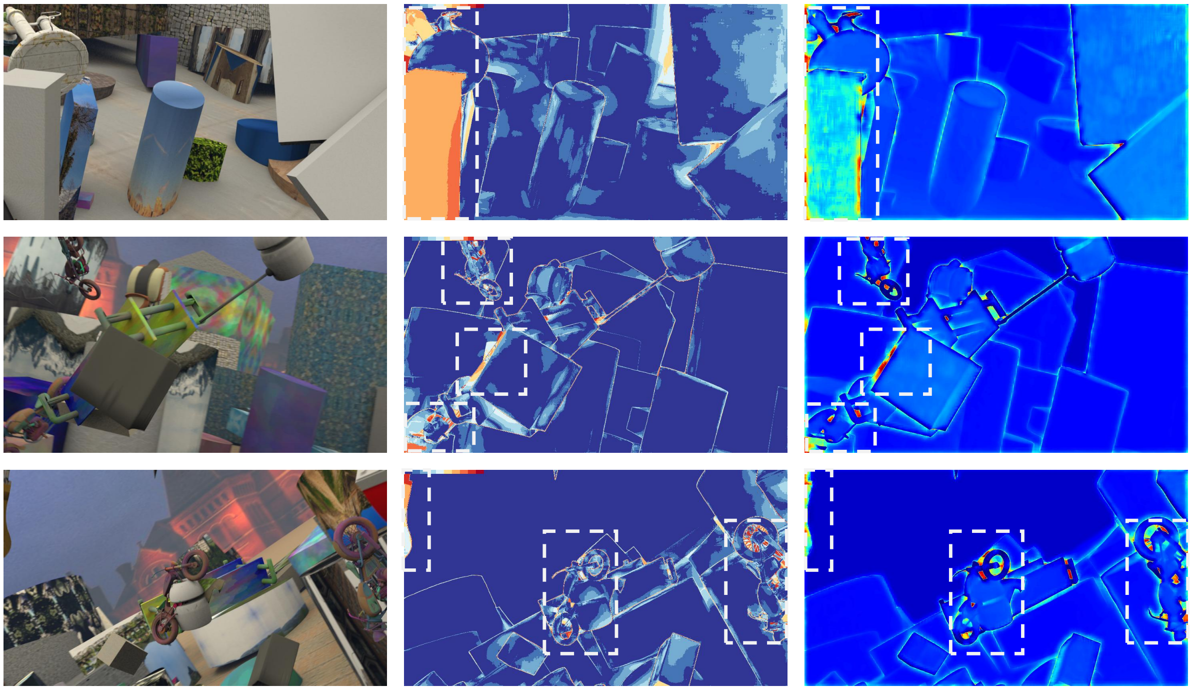

Figure 2. Visualization of the left image (left), error map (center) and uncertainty map (right).

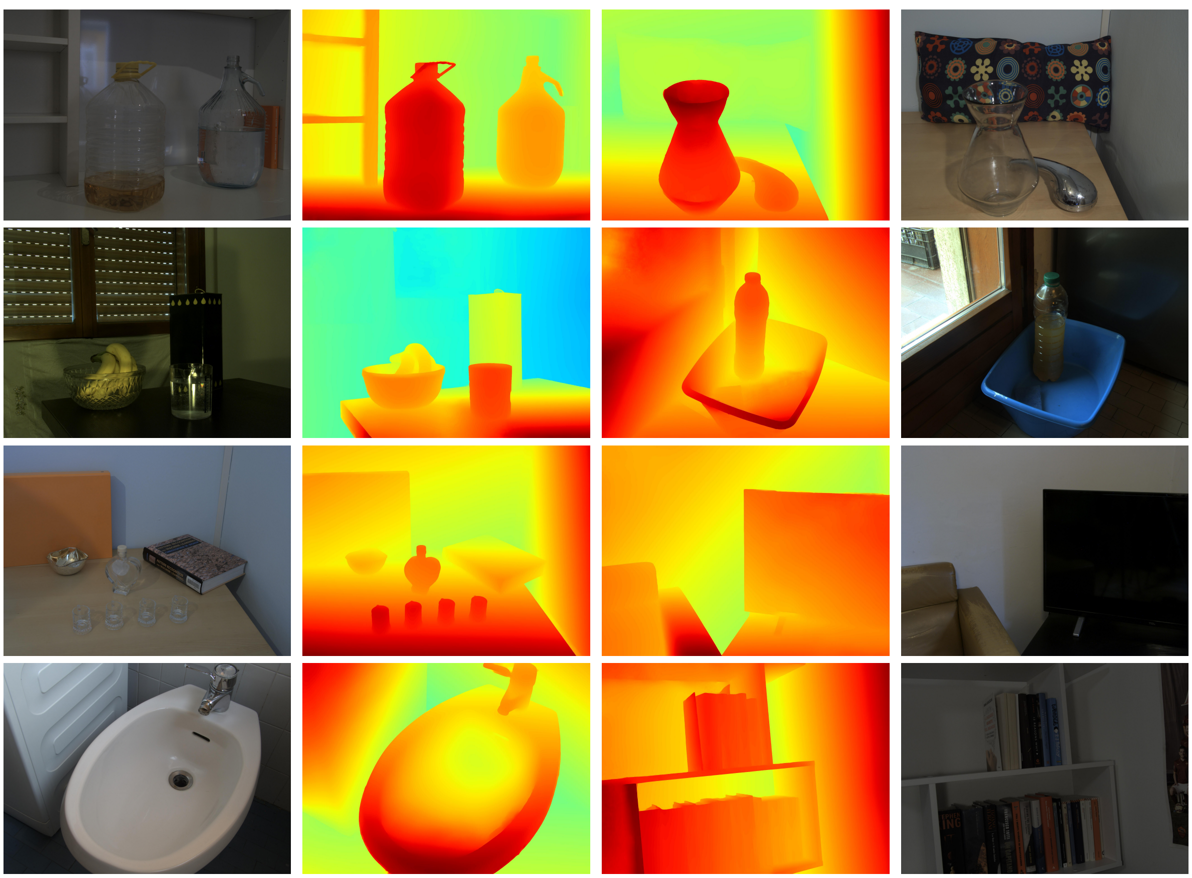

Figure 3. Qualitative results of our method on the booster test set.

Method	Time (s)	All				NonOcc								
		Bad1.0 (%)	Bad2.0 (%)	A50 (px)	A90 (px)	Bad1.0 (%)	Bad2.0 (%)	Bad4.0 (%)	avgerr (px)	rms (px)	A50 (px)	A90 (px)	A95 (px)	A99 (px)
CFNet [8]	0.69	26.2	16.1	0.53	8.37	19.6	10.1	6.49	3.49	15.4	0.48	2.23	16.4	77.6
HITNet [10]	0.14	20.7	12.8	0.45	3.92	13.3	6.46	3.81	1.71	9.97	0.40	2.32	4.26	30.2
HSMNet [12]	0.51	31.2	16.5	0.62	4.26	24.6	10.2	4.82	2.07	10.3	0.56	2.12	4.32	39.2
DeepPruner [2]	0.13	57.1	36.4	1.41	17.9	52.3	30.1	15.9	4.80	14.7	1.17	10.4	23.6	67.7
CREStereo [3]	3.55	14.0	8.13	0.38	1.63	8.25	3.71	2.04	1.15	7.70	0.26	0.92	1.58	22.9
RAFT-stereo [4]	11.6	15.1	9.37	0.37	2.24	9.37	4.74	2.75	1.27	8.41	0.28	1.10	2.29	21.7
PCVNet(ours)	0.18	25.5	13.6	0.54	3.06	19.5	8.19	3.71	1.53	8.71	0.49	1.75	3.08	24.1

Table 2. The supplementary results on Middlebury 2014 dataset [7].

method performs better than other methods with a runtime smaller than 100 ms except HITNet [10]. Nevertheless, it is important to note that our method surpasses HITNet across all pixel areas, as detailed in Table 2 of the main text, which reveals that our method can handle the occlusion regions better than HITNet.

4.3. Middlebury

Table 2 is the supplementary results of Table 3 in the main text. The $Badx$ means the $x$px-error rate and $avgerr$ and rms represent the average absolute error and the root-mean-square error, respectively. The metric Ax in the table is the x-percent error quantile in pixels.

Our method performs well on the $avgerr$, rms, $A90$, $A95$ and $A99$, but get slightly less impressive results on the $Bad1.0$ and $Bad2.0$. It reveals that our model leads to fewer distinct outliers but could be better at fine-grained matching with high-resolution inputs. This might be because of the sparsity of the sampling and can be improved by increasing the number of sample points or slightly increasing the number of iterations to allow the variance to thoroughly converge.

4.4. Booster

Figure 3 shows the visualization of our disparity map on the booster dataset [6]. Our method exhibits remarkable performance in texture-less regions, as well as in challenging areas with reflections and transparency.

References

[1] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[2] Shivam Duggal, Shenlong Wang, Wei-Chiu Ma, Rui Hu, and Raquel Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4384–4393, 2019.

[3] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16263–16272, 2022.

[4] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*, pages 218–227. IEEE, 2021.

[5] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015.

[6] Pierluigi Zama Ramirez, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Open challenges in deep stereo: the booster dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21168–21178, 2022.

[7] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR))*, pages 31–42, 2014.

[8] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13906–13915, 2021.

[9] Zhelun Shen, Yuchao Dai, Xibin Song, Zhibo Rao, Dingfu Zhou, and Liangjun Zhang. Pcw-net: Pyramid combination and warping cost volume for stereo matching. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 280–297. Springer, 2022.

[10] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14362–14372, 2021.

[11] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020.

[12] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2019.

[13] Chengtang Yao, Yunde Jia, Huijun Di, Pengxiang Li, and Yuwei Wu. A decomposition model for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6091–6100, 2021.