CFNet: Cascade and Fused Cost Volume for Robust Stereo Matching Supplementary Material

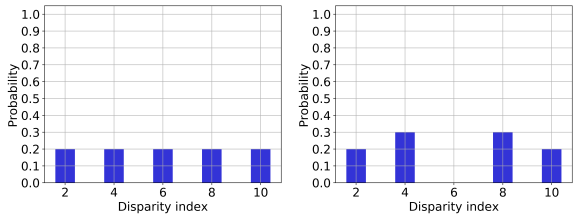
Zhelun Shen[1,2], Yuchao Dai[1]*, Zhibo Rao[1]
[1]Northwestern Polytechnical University, Xi'an, China [2]Peking University, Beijing, China
shenzhelun@pku.edu.cn, daiyuchao@nwpu.edu.cn, raoxi36@foxmail.com

1. Discussion

Differentiability of Cascade Cost Volume: We use Eq.3 (main paper) to generate the next stage's disparity map, which is denoted as:

$$\widehat{d^i} = \sum_{\forall d^i} d \times \sigma(-c_d^i) \tag{1}$$

where the hypothetical disparity index d^i is adjusted by uncertainty estimation. Compared with the Eq.2 (main paper) used in GC-Net [2], we replaced the dense disparity index (hypothesis plane interval equals to 1) with the sparse one. Such an operation does not influence the differentiability of Eq.3.



$\hat{d} = 6.0, U = 8.0$ $\hat{d} = 6.0, U = 8.8$

Figure 1. Some corner cases of multi-modal distribution. Expected value (ground truth) is 6px. The disparity searching range is from 2 to 10 with 5 hypothesis planes.

Limitation of Uncertainty Estimation: Although our uncertainty estimation can generate a reasonable uncertainty map to evaluate the confidence of disparity estimation, limitations still exist. As shown in Figure 1, we notice that some special multi-modal distribution can achieve accurate estimation with high uncertainty. However, these cases appear with little probability and most of the multi-modal distribution leads to inaccurate estimation according to our visualization in Figure 8 of the main paper. In addition, although these cases can achieve accurate estimation, their disparity probability distribution is unreasonable and will hinder overall performance if this special case is underconstrained [7]. Thus, we tend to set a large disparity search range to push the network to generate a predominantly unimodal disparity probability distribution at the next stage. In future work, we plan to distinguish and constraint this special case by better design our formula of uncertainty estimation.

Method	Stage	KITTI D1_all	Middlebury bad 2.0	ETH3D bad 1.0
Casstereo	**stage2**	3.44	43.71	9.39
	stage1	1.78	34.81	4.44
CFNet(ours)	**stage3**	3.54	40.99	8.92
	stage2	2.15	27.04	5.33
	stage1	1.71	22.27	3.57

Table 1. Comparison of each stage's performance between CFNet and Casstereo on KITTI, Middlebury, and ETH3D validation set. The initial disparity estimation result of each method is bolded and underlined. As shown, the stage three (1/8 of the original input image resolution) result of our CFNet can surpass the stage two result of Casstereo (1/4 of the original input image resolution) on two datasets.

2. Additional Comparison with Casstereo

In this section, we first give a more specific mathematical definition to show the difference between our method and Casstereo [1] when estimating the initial disparity. Then we employ two comparative experiments to further show the superiority of our method.

Specifically, the first stage disparity searching index of Casstereo is defined as:

$$\begin{gathered} d^i = 0 + n(\tfrac{D_{\max}}{2^i} - 1)/(N^i - 1) \\ n \in \{0, 1, 2.....N^i - 1\}, i = 2 \end{gathered} \tag{2}$$

where N^i is the number of hypothesis planes at stage i. Instead, the first stage disparity searching index of CFNet is defined as:

$$\begin{gathered} d^i = 0 + n \\ n \in \{0, 1, 2.....\tfrac{D_{\max}}{2^i} - 1\}, i \in (3, 4, 5) \end{gathered} \tag{3}$$

*Corresponding author

Method	stage	Joint Generalization coverage ratio(%)			Cross-domain Generalization coverage ratio(%)		
		KITTI	Middlebury	ETH3D	KITTI	Middlebury	ETH3D
Casstereo	stage2	100	99.69	100	100	99.15	100
	stage1	**99.75**	92.70	99.54	97.02	86.94	99.74
CFNet(ours)	stage3	100	99.92	100	100	99.50	100
	stage2	99.99	99.19	99.73	99.30	97.48	99.98
	stage1	99.70	**97.24**	**99.56**	**98.23**	**93.77**	**99.83**

Table 2. Disparity search ranges setting evaluation in terms of joint generalization and cross-domain generalization. Coverage radio denotes the percentages of disparity search range that cover the ground truth depth. The final disparity estimation result of each method is bolded and underlined.

Note that $(\frac{D_{\max}}{2^i} - 1)/(N^i - 1) \geq 4$ in the stage two of Casstereo. That is we employ multi-scale small-resolution dense cost volume fusion to replace single higher resolution sparse cost volume for initial disparity estimation.

Next, to make a fair comparison, we use the same training strategy to pretrain Casstereo on Scene Flow dataset and finetune it on KITTI 2015, Middlebury, and ETH3D datasets (fixing the disparity search range to 256) and compare with our method. As shown in Table 1, the stage three result (1/8 of the original input image resolution) of our method can even outperform the stage two result (1/4 of the original input image resolution) of Casstereo on two datasets, which further supports our claim that multi-scale small-resolution dense cost volumes fusion can generate a more accurate initial disparity estimation than single higher resolution sparse cost volume.

In addition, as both methods select to iteratively narrow down the disparity space and improve the cost volume resolution, the ratio of the pixels whose generated disparity search range cover the ground truth disparity is an essential indicator to evaluate the reasonability of current disparity search range setting. As shown in Table 2, we evaluate the coverage ratio in two terms of generalization. Comparing with Casstereo, our method can better adjust the disparity search range according to different datasets, especially on the Middlebury. The gap is larger when generalizing to unseen scenes.

3. Details of the Architecture

Table 3 presents the details of our pyramid feature extraction. We employ it to extract multi-scale image features. The final output of each scale is bolded and underlined.

4. More Results

4.1. Error Map vs Uncertainty Map

We give more comparison between each stage's error map and uncertainty map in Figure 2. As shown, the error map is highly correlated with the uncertainty map on all three real datasets. In addition, the higher resolution uncertainty map can better identify the error regions, which further emphasizes the effectiveness of our uncertainty estimation in evaluating the pixel-level confidence of disparity estimation.

4.2. Generalization Results on Different Datasets

In this section, we give more results about the two kinds of generalization we defined in the main paper. As shown in Figure 3, our method can perform well on all three datasets when trained on the same training images and tested on three datasets with single model parameters and hyper-parameters.

Cross-domain generalization is shown in Figure 4. When only trained on synthetic datasets and generalized to real-world datasets, our method can significantly surpass dataset-specific methods [6, 1].

4.3. Finetuning Results on KITTI Dataset

We visualize the results of our fine-tuned CFNet on KITTI2015 and KITTI 2012 datasets and compare it with some state-of-the-art real-time methods [3, 5] in Figure 5. Our method can generate more extract estimation results in the fence and texture-less regions (see dash boxes in the picture).

References

[1] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2020. 1, 2, 7

[2] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 1

[3] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1959–1968, 2020. 2, 8

[4] Gengshan Yang, Joshua Manela, Michael Happold, and Deva Ramanan. Hierarchical deep stereo matching on high-resolution images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5515–5524, 2019. 6

[5] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation.

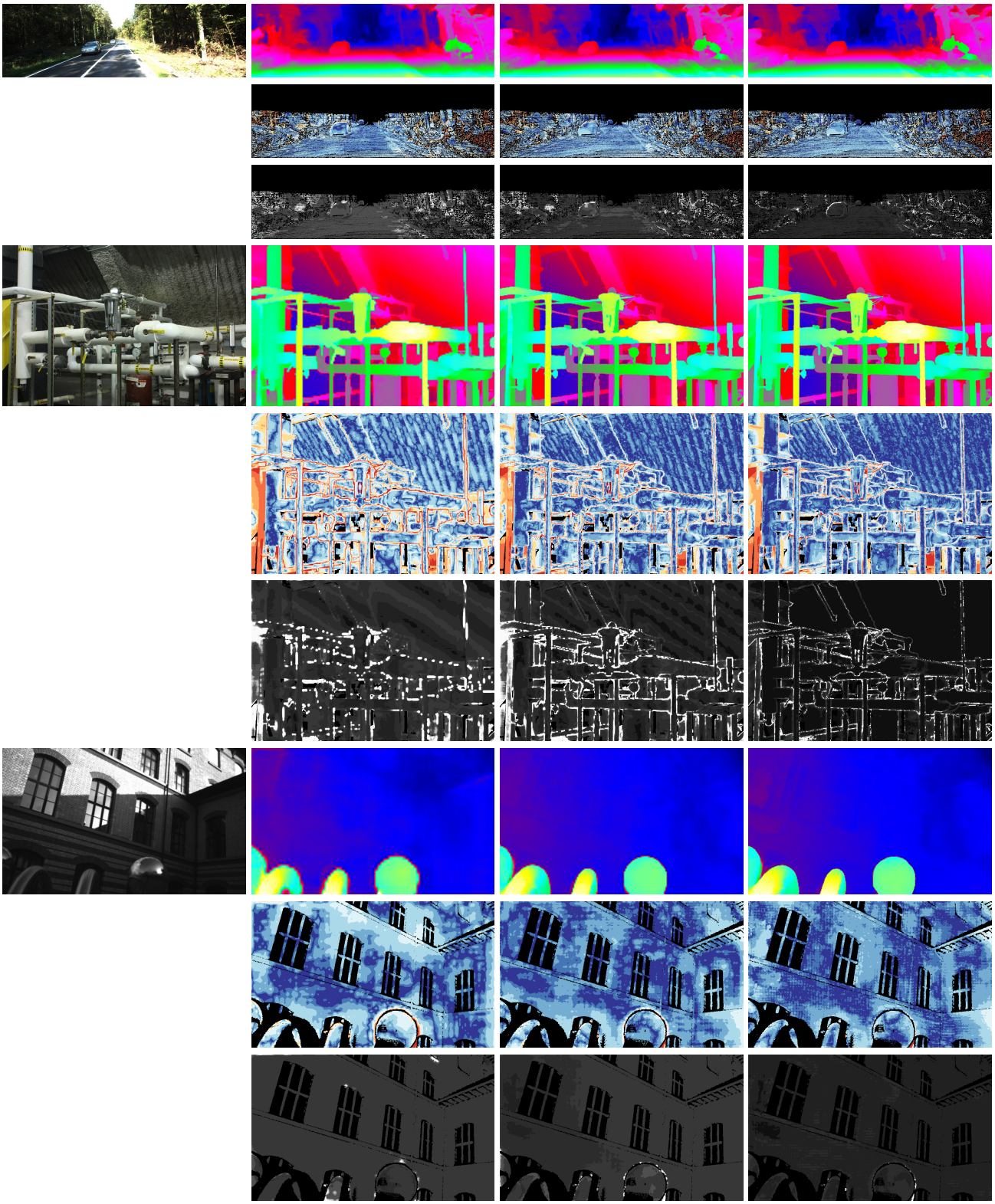
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6044–6053, 2019. 2, 8

[6] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 185–194, 2019. 2, 6, 7

[7] Youmin Zhang, Yimin Chen, Xiao Bai, Suihanjin Yu, Kun Yu, Zhiwei Li, and Kuiyuan Yang. Adaptive unimodal cost volume filtering for deep stereo matching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 12926–12934, 2020. 1

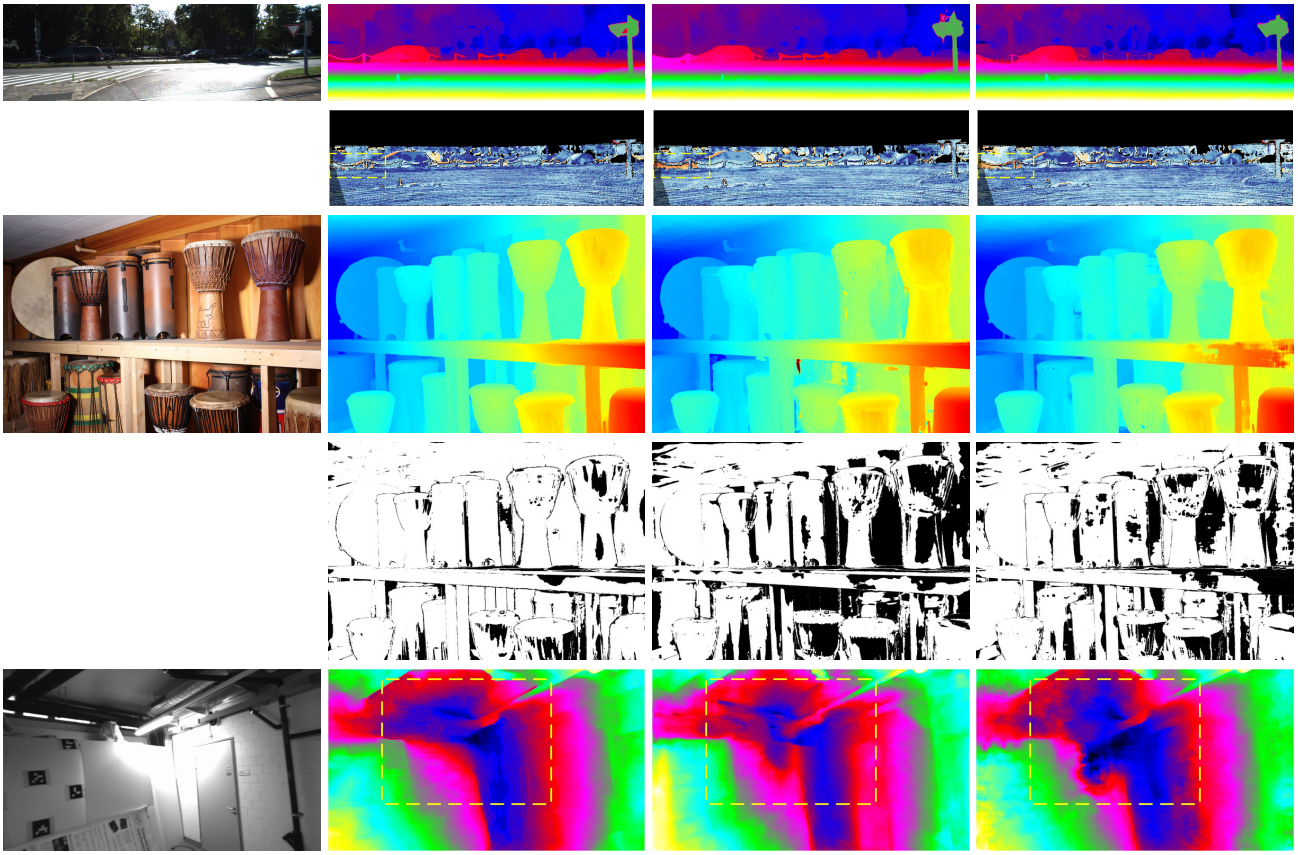
Output	input	Layer Description(k,s,f)	Output dimension
conv0_1	picture	$3 \times 3, 2, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_2	conv0_1	$3 \times 3, 1, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_3	conv0_2	$3 \times 3, 1, 32$	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
conv0_4	conv0_3	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix}$	$\frac{1}{2}H \times \frac{1}{2}W \times 64$
conv0_5	conv0_4	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix}$, stride = 2	$\frac{1}{4}H \times \frac{1}{4}W \times 128$
conv0_6	conv0_5	$\begin{bmatrix} 3 \times 3, 192 \\ 3 \times 3, 192 \end{bmatrix}$, stride = 2	$\frac{1}{8}H \times \frac{1}{8}W \times 192$
conv0_7	conv0_6	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix}$, stride = 2	$\frac{1}{16}H \times \frac{1}{16}W \times 256$
conv0_8	conv0_7	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix}$, stride = 2	$\frac{1}{32}H \times \frac{1}{32}W \times 512$
conv0_9	conv0_8	SPP	$\frac{1}{32}H \times \frac{1}{32}W \times 512$
deconv0_10	conv0_9	Upsample $3 \times 3, 1, 256$	$\frac{1}{16}H \times \frac{1}{16}W \times 256$
deconv0_11	deconv0_10 conv0_7	concat	$\frac{1}{16}H \times \frac{1}{16}W \times 512$
deconv0_12	deconv0_11	$3 \times 3, 1, 256$	$\frac{1}{16}H \times \frac{1}{16}W \times 256$
deconv0_13	deconv0_12	Upsample $3 \times 3, 1, 192$	$\frac{1}{8}H \times \frac{1}{8}W \times 192$
deconv0_14	deconv0_13 conv0_6	concat	$\frac{1}{8}H \times \frac{1}{8}W \times 384$
deconv0_15	deconv0_14	$3 \times 3, 1, 192$	$\frac{1}{8}H \times \frac{1}{8}W \times 192$
deconv0_16	deconv0_15	Upsample $3 \times 3, 1, 128$	$\frac{1}{4}H \times \frac{1}{4}W \times 128$
deconv0_17	deconv0_16 conv0_5	concat	$\frac{1}{4}H \times \frac{1}{4}W \times 256$
deconv0_18	deconv0_17	$3 \times 3, 1, 128$	$\frac{1}{4}H \times \frac{1}{4}W \times 128$
deconv0_19	deconv0_18	Upsample $3 \times 3, 1, 64$	$\frac{1}{2}H \times \frac{1}{2}W \times 64$
deconv0_20	deconv0_19 conv0_4	concat	$\frac{1}{2}H \times \frac{1}{2}W \times 128$
deconv0_21	deconv0_20	$3 \times 3, 1, 64$	$\frac{1}{2}H \times \frac{1}{2}W \times 64$

Table 3. Detailed network structure of our pyramid feature exaction module. Each convolutional layer is followed by the batch normalization and activation function. The final output of each scale is bolded and underlined.



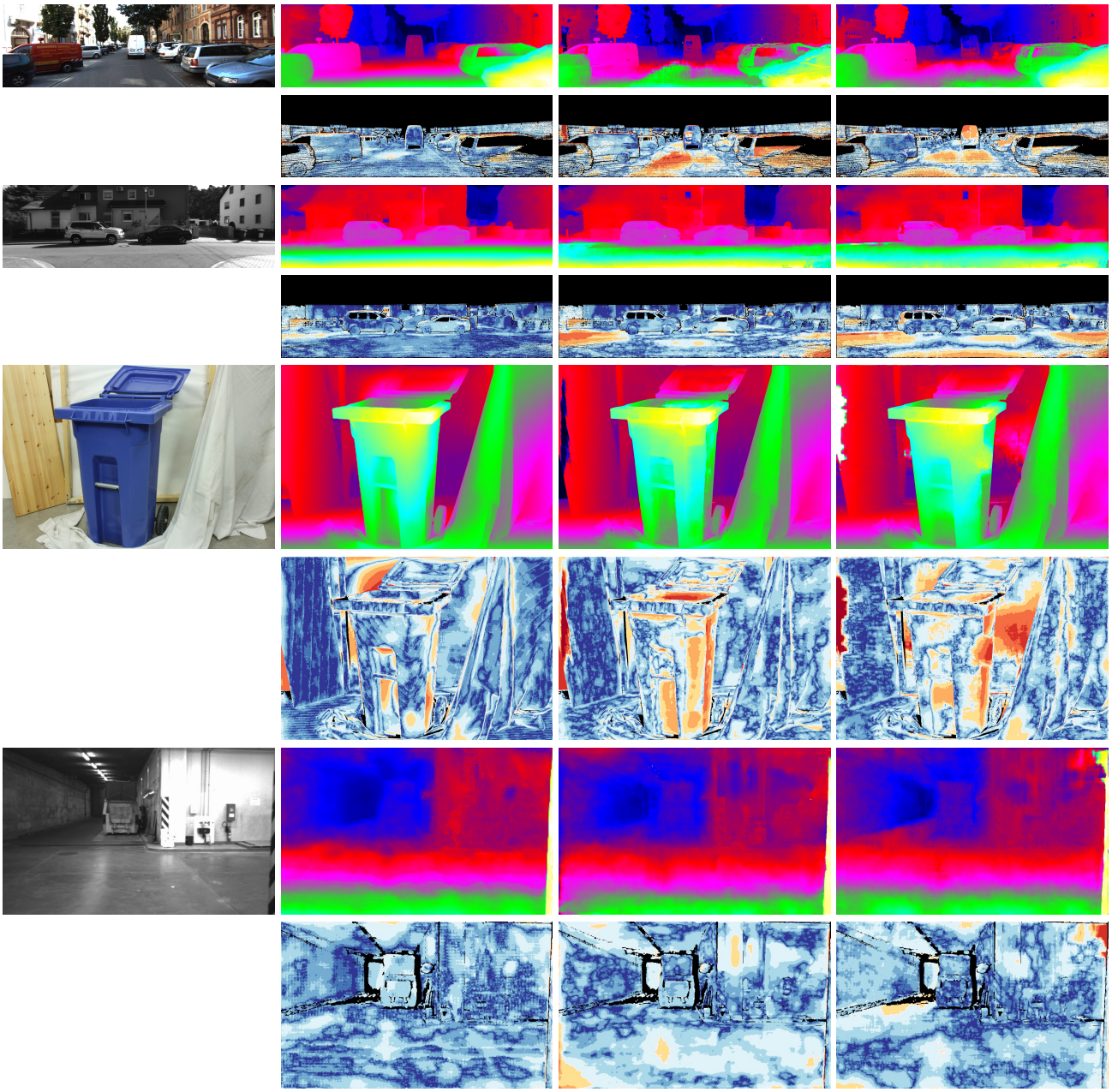
(a) left image (b) stage 3 (c) stage 2 (d) stage 1

Figure 2. Comparison between each stage's error map and uncertainty map on three real datasets (from top to bottom: KITTI2015, Middlebury, and ETH3D). The left panel shows the left input image of the stereo image pair, and for each example, the first row shows the disparity, the second row shows the error map and the third row shows the uncertainty map. Red and white denote large errors and high uncertainty, respectively.



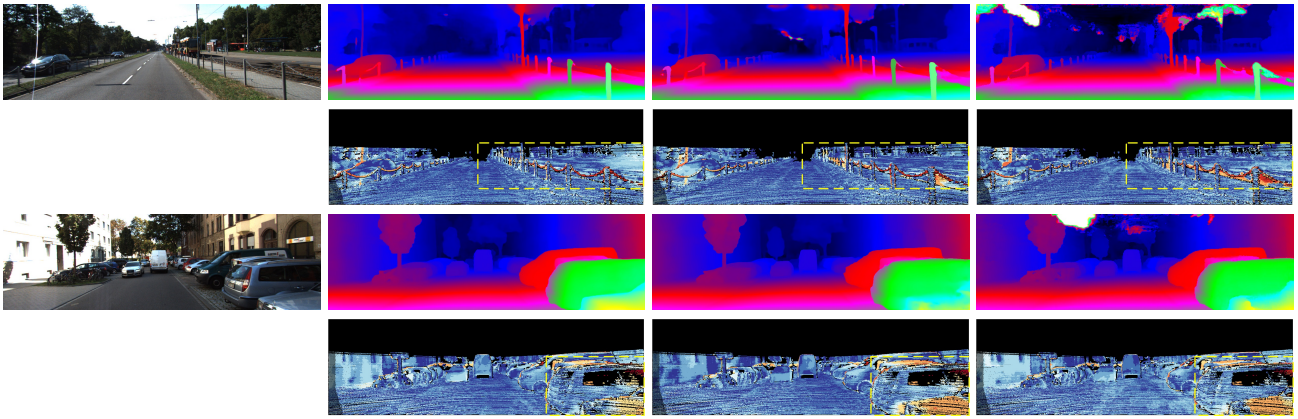
(a) left image (b) CF-Net (c) GANet [6] (d) HSMNet [4]

Figure 3. More visualization of some state-of-the-art methods' generalization ability on three real-world dataset testsets (from top to bottom: KITTI2015, Middlebury, and ETH3D). The left panel shows the left input image of the stereo image pair, and for each example, the first row shows the predicted colorized disparity map and the second row shows the error map (we omit the error map of ETH3D because the evaluation server doesn't provide it). Our CFNet achieves SOTA or near SOTA performance on all three datasets without any adaptation.

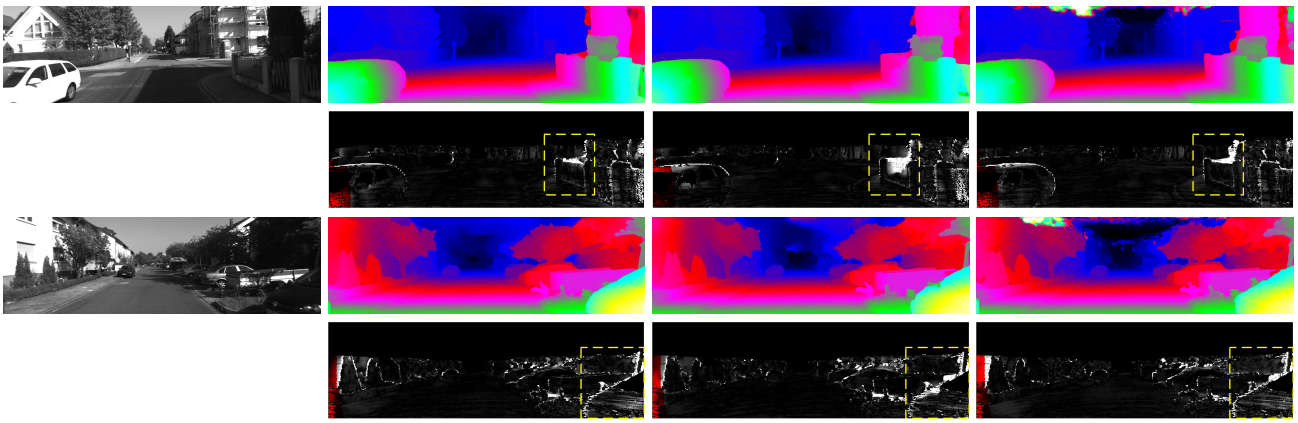


(a) left image (b) CF-Net (c) GANet[6] (d) Casstereo[1]

Figure 4. Unseen scene generalization evaluation on ETH3D, Middlebury, and KITTI training sets (from top to bottom: KITTI2015, KITTI2012, Middlebury, and ETH3D). All methods are only trained on the Scene Flow datatest and tested on full-resolution training images of four real datasets. The left panel shows the left input image of the stereo image pair, and for each example, the first row shows the predicted colorized disparity map and the second row shows the error map.



KITTI2015



KITTI2012

(a) left image (b) CF-Net (c) AANet [3] (d) HD^3 [5]

Figure 5. Visualization results of our finetune model on the KITTI dataset testset.The left panel shows the left input image of the stereo image pair, and for each example, the first row shows the predicted colorized disparity map and the second row shows the error map. Our method can generate more extract estimation results in the fence and texture-less regions.