Supplementary Material for ”WaveletStereo: Learning Wavelet Coefficients of Disparity Map in Stereo Matching”

Menglong Yang, Fangrui Wu and Wei Li
School of Aeronautics and Astronautics, Sichuan University
Chengdu, Sichuan, PR China
mlyang@scu.edu.cn, wufangrui@stu.scu.edu.cn, li.wei@scu.edu.cn

1. Introduction

In this supplementary material, we firstly describe our network configuration in detail. Then we perform an ablation study on the loss function and showcase more visual results of the proposed WaveletStereo, which enable the readers to quantificationally and qualitatively understand the effectiveness of the proposed mechanism.

2. Network configuration

In this paper, we divided the network architecture into three modules, i.e., deep representation, multi-resolution cost volumes and multi-resolution reconstruction. In the remainder of this section, we respectively describe the configuration of each module in detail.

2.1. Deep representation

The deep representation extracts unary features from a pair of stereo images with a shared weight Siamese network, to form a cost volume. As Table 1 shows, there are two downsampling modules in the deep representation, each of which is followed by a densely connected atrous spatial pyramid block. The downsampling is simply performed by using convolutions with $3 \times 3 \times 32$ filters and a stride of 2. The last layer of the deep representation is a convolution operation with $3 \times 3 \times 32$ filters, and the outputs are $\frac{1}{4}H \times \frac{1}{4}W \times 32$ features for the left and right image, respectively. Each convolution is followed by a batch normalization and ReLU activation except the output layer.

2.2. Multi-resolution cost volumes

The module of multi-resolution cost volumes forms several cost volumes with different resolutions from the unary features.

As Table 2 shows, we first construct a cost volume with a resolution of $\frac{1}{4}H \times \frac{1}{4}W$ by concatenating the left unary feature and the right unary features with shifted disparities. Then two downsampling modules are used to form cost volumes with the resolutions of $\frac{1}{8}H \times \frac{1}{8}W$ and $\frac{1}{16}H \times \frac{1}{16}W$, respectively. There is a 3D densely connected atrous spatial pyramid block in the cost volume of each resolution.

2.3. Multi-resolution wavelet reconstruction

The module of multi-resolution wavelet reconstruction maps the multi-resolution cost volumes into the multi-resolution wavelet coefficients and performs the wavelet reconstruction level by level through inverse wavelet transforms.

As Table 3 shows, we iteratively repeat a similar process, i.e., mapping a cost volume to the wavelet coefficients and calculating the low-frequency coefficient of a higher resolution by inverse wavelet transform. To reduce the error accumulation in the progress of multi-resolution reconstruction, we added an operation of edge-aware filtering after each step of inverse wavelet transform, which is similar to the refinement operation proposed in StereoNet [?]. The detailed configuration of the edge-aware filtering is shown in Table 5.

3. Ablation study on loss function

In this paper, we use a loss function containing two terms, i.e.,

$$L_1 = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N} smooth_{L_1} \left(\psi_i^l - \hat{\psi}_i^l \right), \tag{1}$$

and

$$\mathcal{L}_2 = \frac{1}{N} \sum_{i=1}^{N} smooth_{L_1} \left(\hat{d}_i - d_i \right) \tag{2}$$

We performed three experiments to analyze the effectiveness of the two loss terms, as shown in Table 4. We can see the positive effectiveness of the two loss terms, by training the network on Scene Flow with L_1, L_2 and $L_1 + L_2$ in sequence, and comparing the test results on Scene Flow test set.

Table 1. Configuration of deep representation. Each convolutional layer represents a block of convolution, batch normalization and ReLU non-linearity (unless otherwise specified).

	Layer Description	Output Tensor Dim.
	Input image	$H \times W \times 3$
1	3×3 conv, 32 features, stride 2	$\frac{1}{2}H \times \frac{1}{2}W \times 32$
	Densely connected atrous spatial pyramid block	
2-5	from layer 1, repeat 3×3 conv with 4 features and dilated rate of 1, 2, 4 and 8	$\frac{1}{2}H \times \frac{1}{2}W \times 4$
6	concat layers $1-5$	$\frac{1}{2}H \times \frac{1}{2}W \times 48$
7-10	from layer 6, repeat 3×3 conv with 4 features and dilated rate of 1, 2, 4 and 8	$\frac{1}{2}H \times \frac{1}{2}W \times 4$
11	concat layers $6-10$	$\frac{1}{2}H \times \frac{1}{2}W \times 64$
	Densely connected atrous spatial pyramid block	
12	3×3 conv, 32 features, stride 2	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
13-14	from layer 12, repeat 3×3 conv with 8 features and dilated rate of 1 and 2	$\frac{1}{4}H \times \frac{1}{4}W \times 8$
15	concat layers $12-14$	$\frac{1}{4}H \times \frac{1}{4}W \times 48$
16-17	from layer 15, repeat 3×3 conv with 8 features and dilated rate of 1 and 2	$\frac{1}{4}H \times \frac{1}{4}W \times 8$
18	concat layers $15-17$	$\frac{1}{4}H \times \frac{1}{4}W \times 64$
19-20	from layer 18, repeat 3×3 conv with 8 features and dilated rate of 1 and 2	$\frac{1}{4}H \times \frac{1}{4}W \times 8$
21	concat layers $18-20$	$\frac{1}{4}H \times \frac{1}{4}W \times 80$
22-23	from layer 21, repeat 3×3 conv with 8 features and dilated rate of 1 and 2	$\frac{1}{4}H \times \frac{1}{4}W \times 8$
24	concat layers $21-23$	$\frac{1}{4}H \times \frac{1}{4}W \times 96$
25	3×3 conv, 32 features, (no ReLu or BN)	$\frac{1}{4}H \times \frac{1}{4}W \times 32$

Table 2. Configuration of multi-resolution cost volumes. Each convolutional layer represents a block of convolution, batch normalization and ReLU non-linearity (unless otherwise specified).

	Layer Description	Output Tensor Dim.
	Input left and right features	$\frac{1}{4}H \times \frac{1}{4}W \times 32$
	Cost volume 1	
26	concat left feature and shift right features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 32$
27	3D conv, $3 \times 3 \times 3$, 16 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
28	concat layers $26-27$	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 48$
29	3D conv, $3 \times 3 \times 3$, 16 features	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
30	concat layers $28-29$	$\frac{1}{4}D \times \frac{1}{4}H \times \frac{1}{4}W \times 64$
	Cost volume 2	
31	3D conv, $3 \times 3 \times 3$, 32 features, stride 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 32$
32-33	from 31, repeat 3D conv of $3 \times 3 \times 3$ with 8 features and dilated rate of 1 and 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 8$
34	concat layers $31-34$	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 48$
35-36	from 34, repeat 3D conv of $3 \times 3 \times 3$ with 8 features and dilated rate of 1 and 2	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 8$
37	concat layers $34-36$	$\frac{1}{8}D \times \frac{1}{8}H \times \frac{1}{8}W \times 64$
	Cost volume 3	
38	3D conv, $3 \times 3 \times 3$, 32 features, stride 2	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 32$
39-42	from 38, repeat 3D conv of $3 \times 3 \times 3$ with 4 features and dilated rate of 1, 2, 4 and 8	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 4$
43	concat layers $38-42$	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 48$
44-47	from 43, repeat 3D conv of $3 \times 3 \times 3$ with 4 features and dilated rate of 1, 2, 3 and 4	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 4$
48	concat layers $44-47$	$\frac{1}{16}D \times \frac{1}{16}H \times \frac{1}{16}W \times 64$

As metioned in Section 4.3, the proposed WaveletStereo achieves the state-of-the-art performance on Scene Flow dataset, but get an average performance on KITTI. The ground truth of KITTI training set is sparsely labeled, and it is difficult for which to compute the high-frequency wavelet coefficients in too many regions. The loss term L_1 almost does not work on KITTI. From Table 4, we can find out the performance under such a condition similar to KITTI, i.e., the accuracy dramaticly declines when the training is only supervised by L_2. It is not hard to understand the different performances of the proposed algorithm on Scene Flow and KITTI.

4. More Visual Results

Our paper has shown many quantitative analyses of the proposed method, but the visual results are not completely shown for conciseness. As a supplement, we showcase more disparity images.

Table 3. Configuration of multi-resolution wavelet reconstruction. Each convolutional or transposed convolutional layer represents a block of convolution, batch normalization and ReLU non-linearity (unless otherwise specified).

	Layer Description	Output Tensor Dim.
	Level 3	
49	from layer 48, 3D transposed conv, $3 \times 3 \times 3$, 32 features, stride $1 \times 2 \times 2$	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 32$
50	3D conv, $3 \times 3 \times 3$, 32 features	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 32$
51	3D conv, $3 \times 3 \times 3$, 32 features	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 32$
52	add features of layer 49 and 51 (residual connection)	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 32$
53	3D conv, $3 \times 3 \times 3$, 1 feature	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W$
54	soft argmin, low-frequency wavelet coeffient	$\frac{1}{8}H \times \frac{1}{8}W$
55	from layer 52, 3D conv, $3 \times 3 \times 3$, 2 features	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2$
56	softmax along the axis of depth	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W \times 2$
57	subtraction between the two features of layer 56	$\frac{1}{16}D \times \frac{1}{8}H \times \frac{1}{8}W$
58	weighted sum, horizontal high frequency coefficient	$\frac{1}{8}H \times \frac{1}{8}W$
59-62	from layer 52, repeat layers 55-58, vertical high frequency coefficient	$\frac{1}{8}H \times \frac{1}{8}W$
63-66	from layer 52, repeat layers 55-58, diagonal high frequency coefficient	$\frac{1}{8}H \times \frac{1}{8}W$
67	from layers 54, 58, 62 and 66, inverse wavelet transform	$\frac{1}{4}H \times \frac{1}{4}W$
	Level 2	
68	edge-aware filtering, refined low-frequency coefficient	$\frac{1}{4}H \times \frac{1}{4}W$
69	from layer 37, 3D transposed conv, $3 \times 3 \times 3$, 16 features, stride $1 \times 2 \times 2$	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
70	3D conv, $3 \times 3 \times 3$, 16 features	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
71	3D conv, $3 \times 3 \times 3$, 16 features	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
72	add features of layer 69 and 71 (residual connection)	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 16$
73	3D conv, $3 \times 3 \times 3$, 2 features	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2$
74	softmax along the axis of depth	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W \times 2$
75	subtraction between the two features of layer 74	$\frac{1}{8}D \times \frac{1}{4}H \times \frac{1}{4}W$
76	weighted sum, horizontal high frequency coefficient	$\frac{1}{4}H \times \frac{1}{4}W$
77-80	from layer 72, repeat layers 73-76, vertical high frequency coefficient	$\frac{1}{4}H \times \frac{1}{4}W$
81-84	from layer 72, repeat layers 73-76, diagonal high frequency coefficient	$\frac{1}{4}H \times \frac{1}{4}W$
85	from layers 68, 76, 80 and 84, inverse wavelet transform	$\frac{1}{2}H \times \frac{1}{2}W$
	Level 1	
86	edge-aware filtering, refined low-frequency coefficient	$\frac{1}{2}H \times \frac{1}{2}W$
87	from layer 30, 3D transposed conv, $3 \times 3 \times 3$, 8 features, stride $1 \times 2 \times 2$	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 16$
88	3D conv, $3 \times 3 \times 3$, 8 features	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 8$
89	3D conv, $3 \times 3 \times 3$, 8 features	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 8$
90	add features of layer 87 and 89 (residual connection)	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 8$
91	3D conv, $3 \times 3 \times 3$, 2 features	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 2$
92	softmax along the axis of depth	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 2$
93	subtraction between the two features of layer 92	$\frac{1}{4}D \times \frac{1}{2}H \times \frac{1}{2}W \times 1$
94	weighted sum, horizontal high frequency coefficient	$\frac{1}{2}H \times \frac{1}{2}W$
95-98	from layer 90, repeat layers 91-94, vertical high frequency coefficient	$\frac{1}{2}H \times \frac{1}{2}W$
98-102	from layer 90, repeat layers 91-94, diagonal high frequency coefficient	$\frac{1}{2}H \times \frac{1}{2}W$
103	from layers 86, 94, 98 and 102, inverse wavelet transform, disparity	$H \times W$
104	edge-aware filtering, refined disparity	$H \times W$

Table 4. Results on the Scene Flow dataset. We compare different loss functions to justify the effectiveness of our design choices.

Model	> 3 px (%)	EPE
L_1 only	5.47	1.02
L_2 only	4.71	0.97
$L_1 + L_2$	4.13	0.84

Fig. 1 and Fig. 2 show the visual results of the Scene Flow dataset [?] and KITTI [?, ?] dataset, respectively. More specfically, we demonstrate the effectiveness of the proposed mechanism in Fig.1. We firstly removed all high-frequency predictors and obtained the results, denoted as LF only in Fig. 1, i.e., the second and third columns. The last two columns are the results of the full model. From the comparison between the two results, one can see that the effect of the proposed high-frequency predictor in decreasing the errors not only in the regions of thin surfaces and sharp edges, but in the slant textureless surfaces.

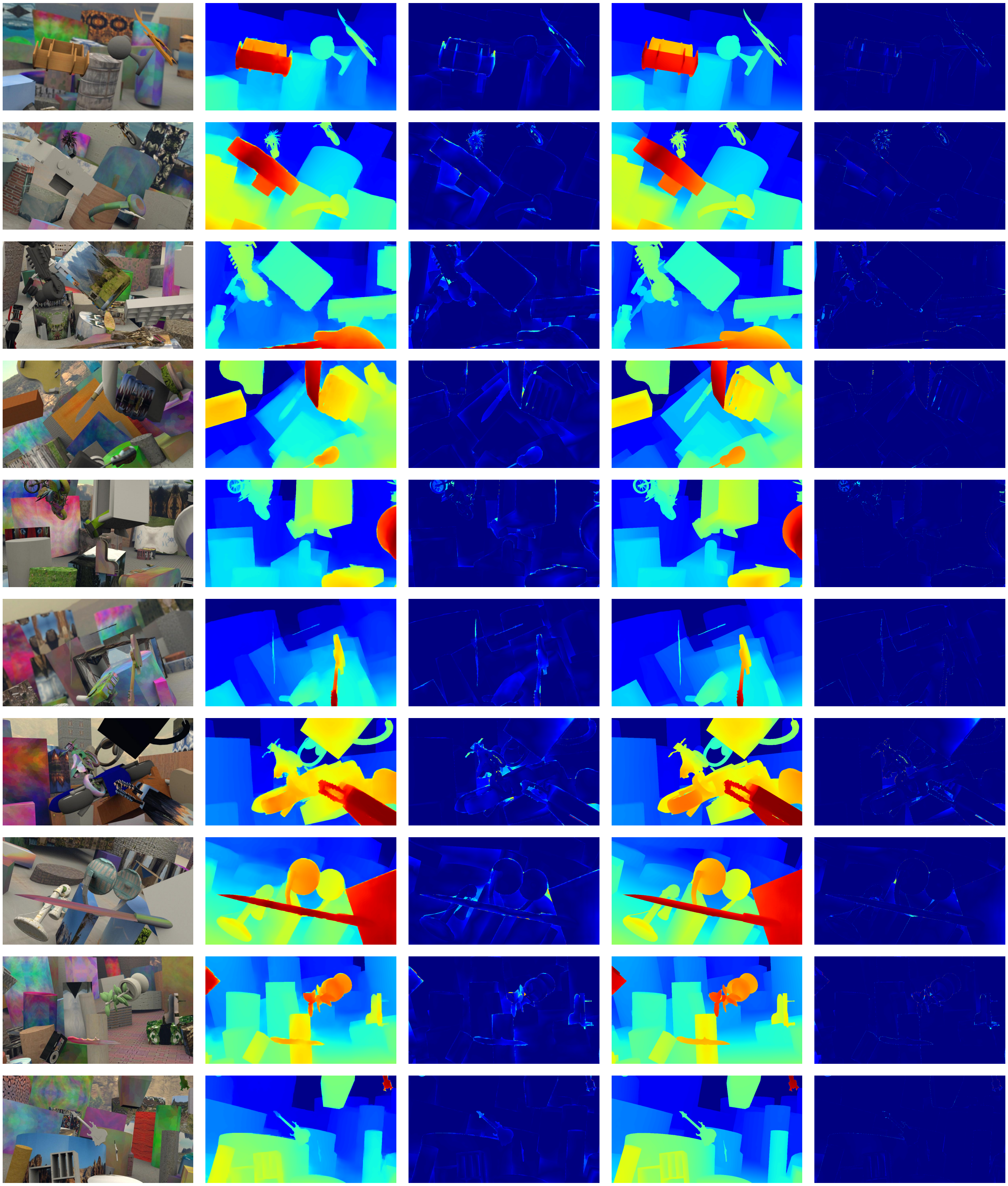
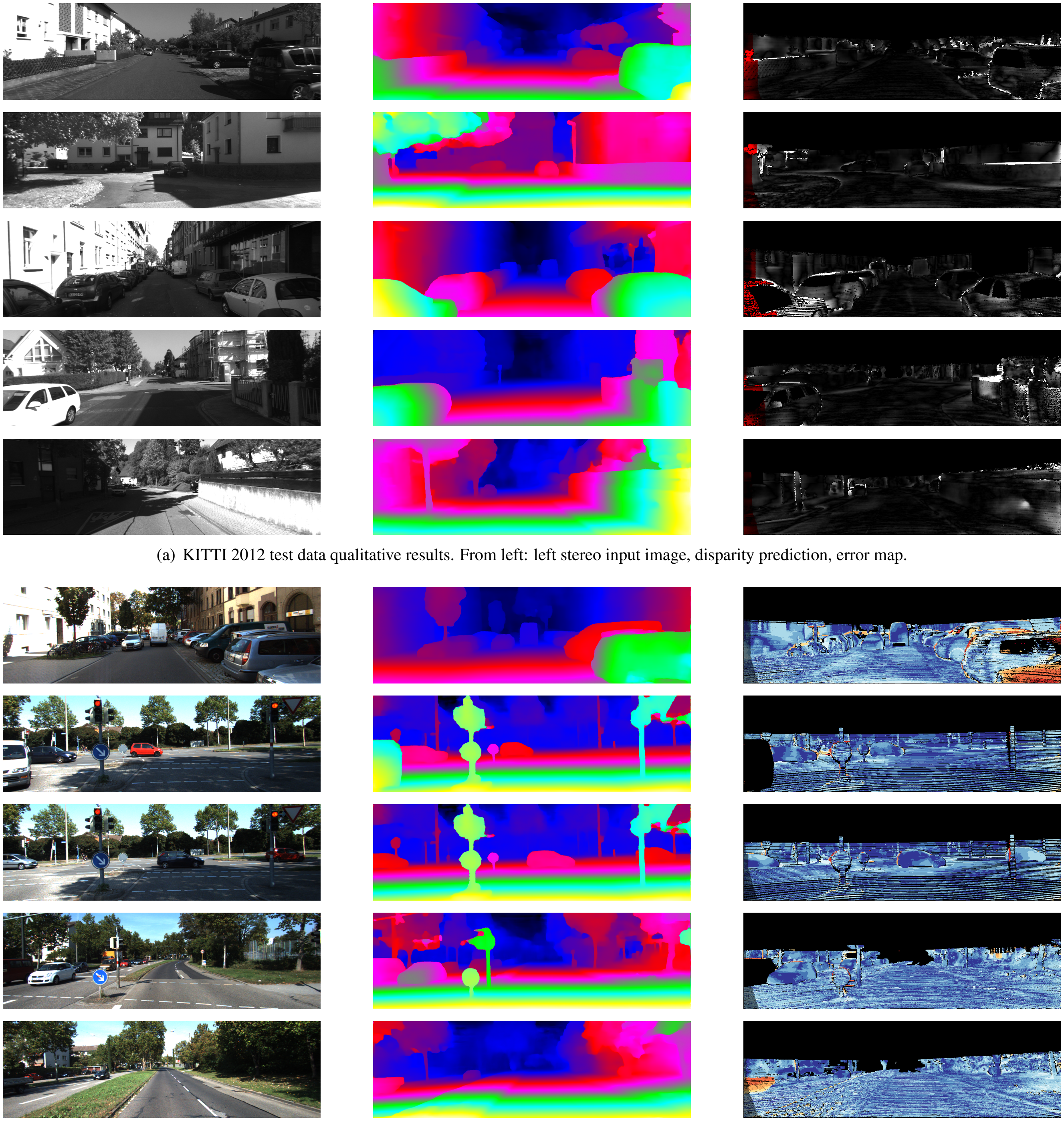


Figure 1. Qualitative results of Scene Flow test set. From left to right: left stereo input image, predicted disparity (LF only), the errors (LF only), predicted disparity of the full model, and the errors of the full model.



(a) KITTI 2012 test data qualitative results. From left: left stereo input image, disparity prediction, error map.

(b) KITTI 2015 test data qualitative results. From left: left stereo input image, disparity prediction, error map.

Figure 2. KITTI test data qualitative results.

Table 5. Configuration of edge-aware filtering architecture. Each convolutional or transposed convolutional layer represents a block of convolution, batch normalization and ReLU non-linearity (unless otherwise specified).

	Layer Description	Output Tensor Dim.
	Input Resized image	$h \times w \times 3$
	Initial low-frequency coefficient	$h \times w$
	concat input resized image and initial low-frequency coefficient	$h \times w \times 4$
1	3×3 conv, 32 features	$h \times w \times 32$
2	3×3 conv, 32 features with dilated rate of 1	$h \times w \times 32$
3	3×3 conv, 32 features with dilated rate of 1	$h \times w \times 32$
4	add features of layer 2 and 4 (residual connection)	$h \times w \times 32$
5-7	from layer 1, repeat layers 2-4 with dilated rate of 2	$h \times w \times 32$
8-10	from layer 7, repeat layers 5-7 with dilated rate of 4	$h \times w \times 32$
11-13	from layer 10, repeat layers 8-10 with dilated rate of 8	$h \times w \times 32$
14-16	from layer 13, repeat layers 11-13 with dilated rate of 1	$h \times w \times 32$
17-19	from layer 16, repeat layers 14-16 with dilated rate of 1	$h \times w \times 32$
20	3×3 conv, 1 feature	$h \times w \times 1$
21	add feature of layer 20 and the initial low-frequency coefficient (residual connection)	$h \times w$