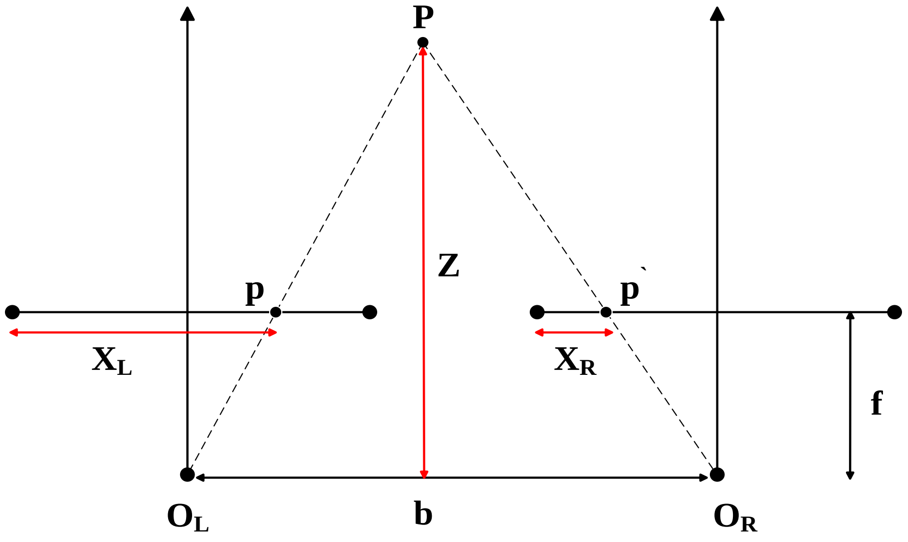
天津理工大学
TIANJIN UNIVRESITY OF TECHNOLOGY

ICPR 2020

Attention Stereo Matching Network

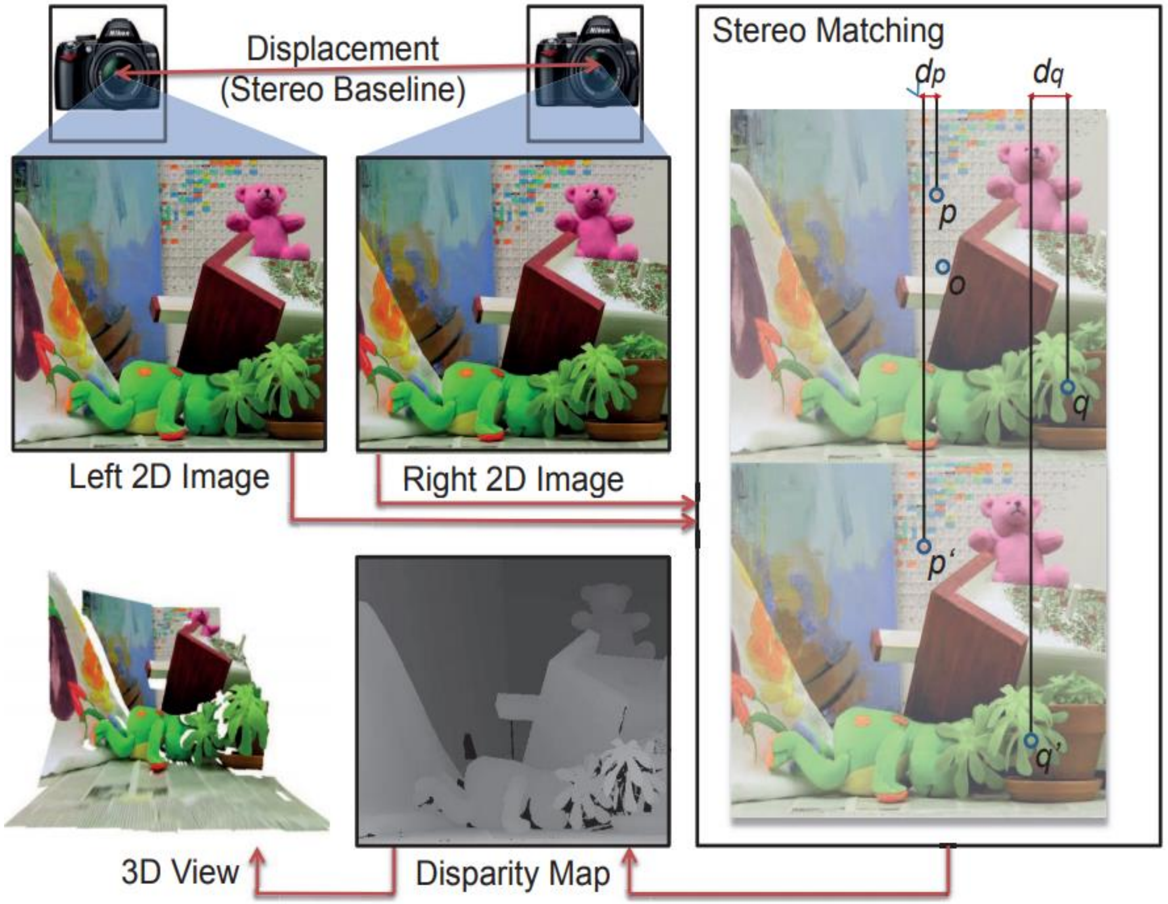
Doudou Zhang[1,*], Jing Cai[1,*], Yanbing Xue[1,†], Zan Gao[2] and Hua Zhang[3]
[1]School of Computer Science and Engineering
Tianjin University of Technology, Tianjin 300384, China
[2]Institute of AI, Shandong Computer Science Center
National Supercomputer Center in Jinan
Qilu University of Technology, Shandong 250101, China
[3]Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China

- **Basic principle of stereo matching**

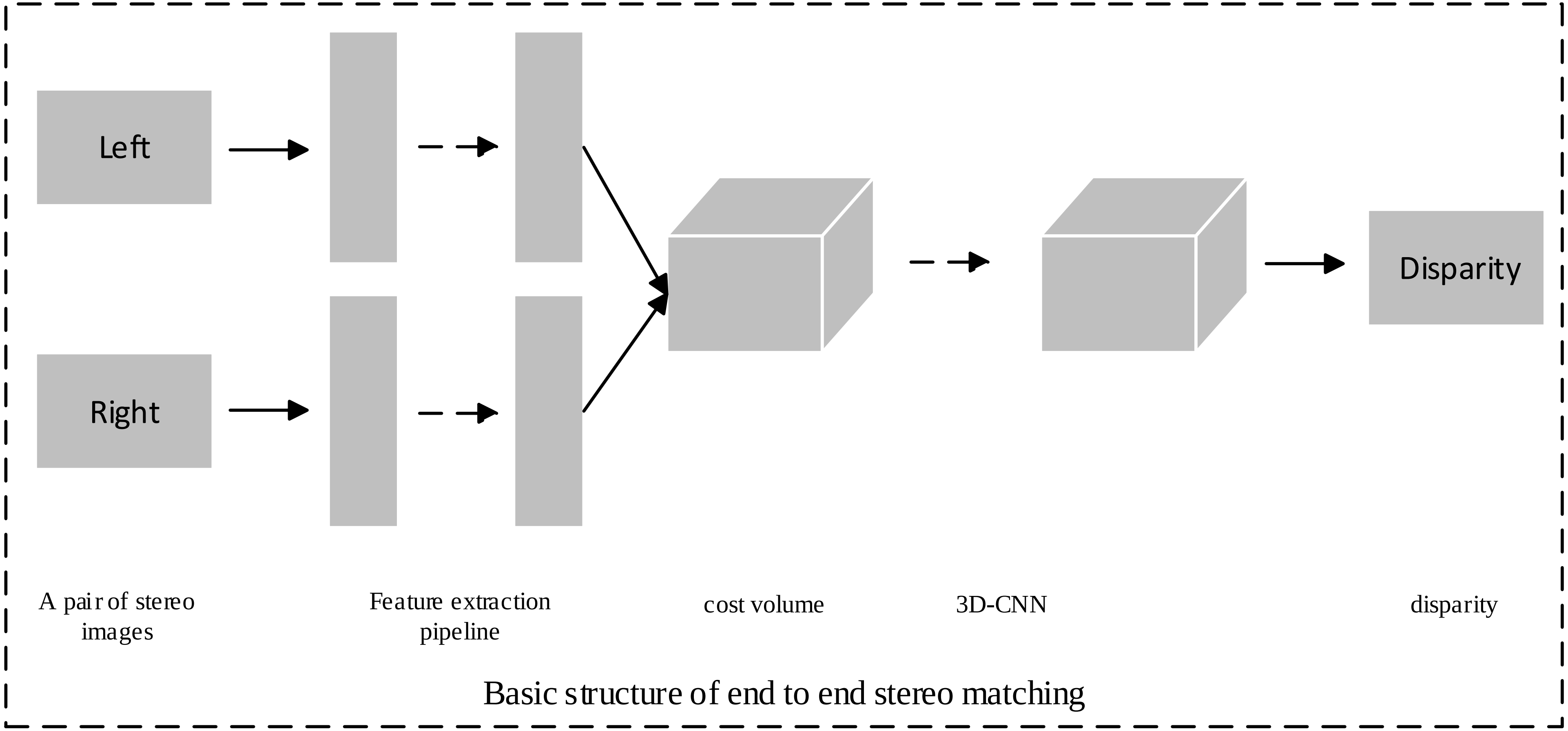


$$\frac{b}{Z}=\frac{b-\left[\frac{L}{2}-\left(L-X_L\right)\right]-\left(\frac{L}{2}-X_R\right)}{Z-f}=\frac{b+X_R-X_L}{Z-f}$$

$$Z=\frac{bf}{X_L-X_R}=\frac{bf}{d}$$



- **Basic structure of end-to-end stereo matching**



Basic structure of end to end stereo matching

Challenge and Related work

- How to improve the matching accuracy in slender structures and textureless regions.
- How to balance the accuracy and the speed.

- The accuracy of the algorithm based on traditional stereo matching can not meet the requirements.
- In recent years, the algorithm based on end-to-end using multi-scale 3D convolution has a large amount of calculation and slow calculation speed.
- Some stereo matching networks based on multi-stage end-to-end deep learning can not balance the relationship between precision and speed.
- There are few application models that can achieve the balance of speed and precision.

● Challenges and Our Corresponding Methods

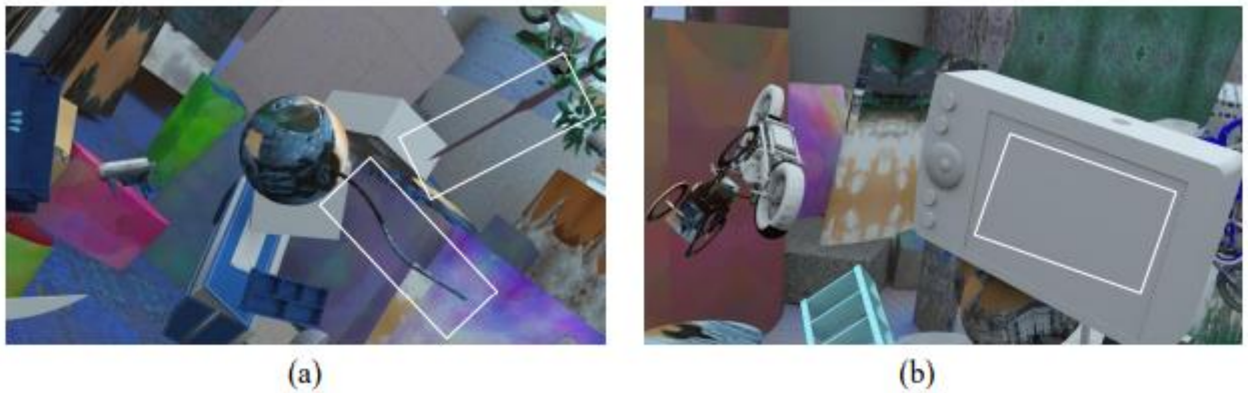
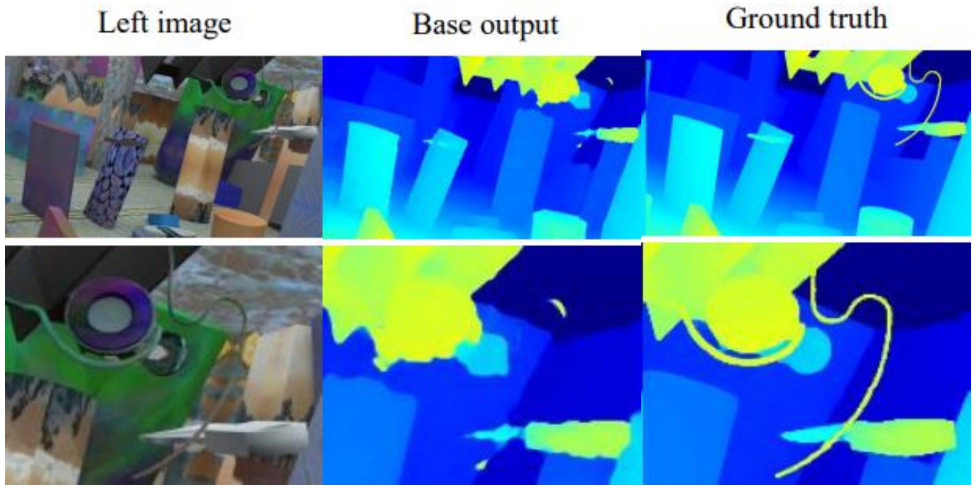


Fig. 1. Challenging regions in Stereo Matching include (a) the slender structures and (b) the textureless regions.



Despite great progress, previous stereo matching algorithms still lack the ability to match textureless regions and slender structure areas. To tackle this problem, we propose ASMNet, an attention stereo matching network. Attention module and disparity refinement module are constructed in the ASMNet. The attention module can improve correlation information between two images by channels and spatial attention.The feature-guided disparity refinement module learns more geometry information in different feature levels to refine the coarse prediction resolution constantly. The proposed approach was evaluated on several benchmark datasets. Experiments show that the proposed method achieves competitive results on KITTI and Scene-Flow datasets while running in real-time at 14ms.

Our Method

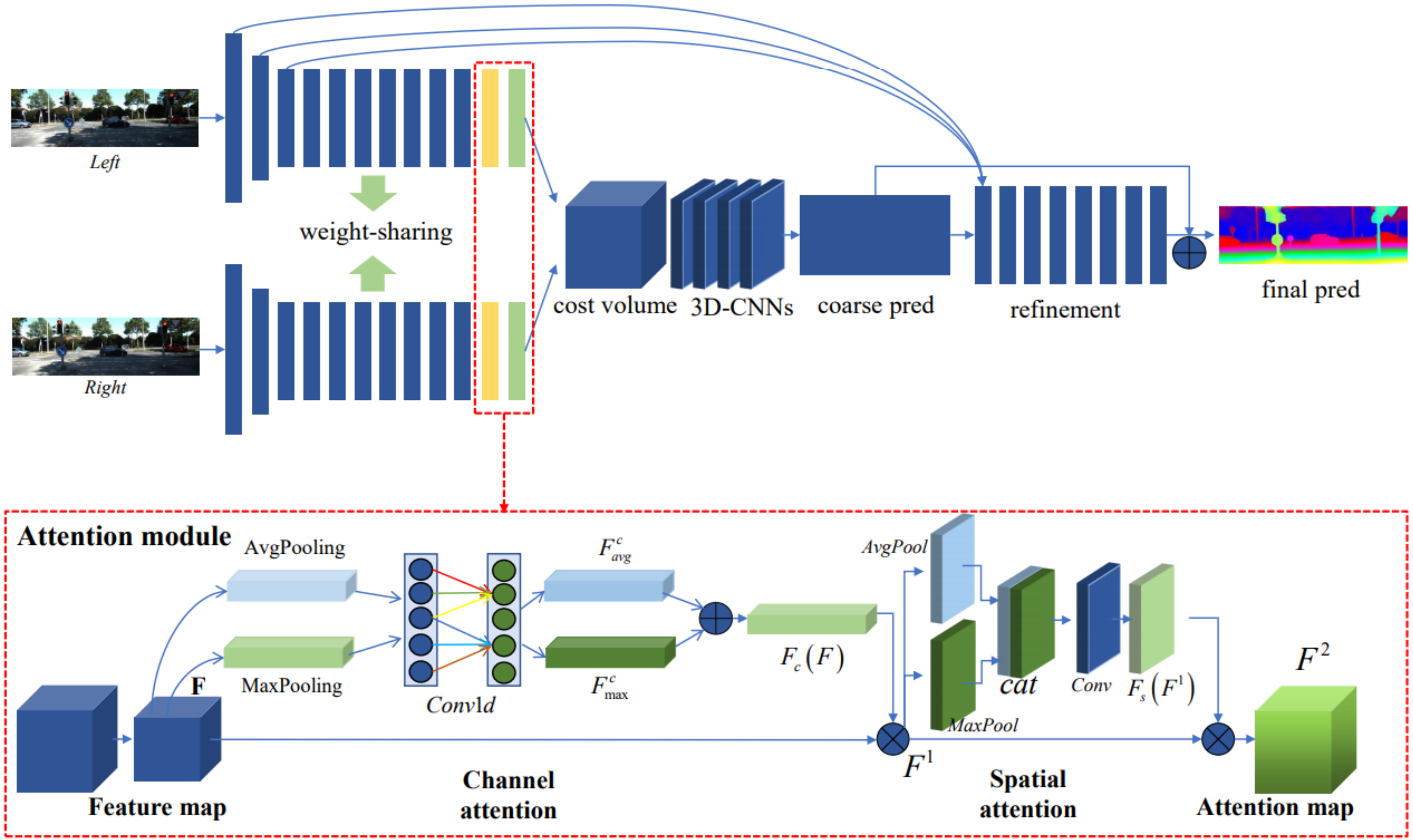


Fig. 2. Architecture overview of proposed ASMNet. A pair of rectified stereo images are fed to a weight-sharing pipelines and down-sample three times rapidly. Then the feature extraction function is provided through a series of residual structures. The feature maps pass through the attention module to get the attention maps. The cost volume is constructed by sliding subtraction of left and right features maps, and the coarse resolution initial disparity map is obtained by a small amount of 3D convolution and disparity regression, then the final prediction result is obtained by refining the coarse disparity.

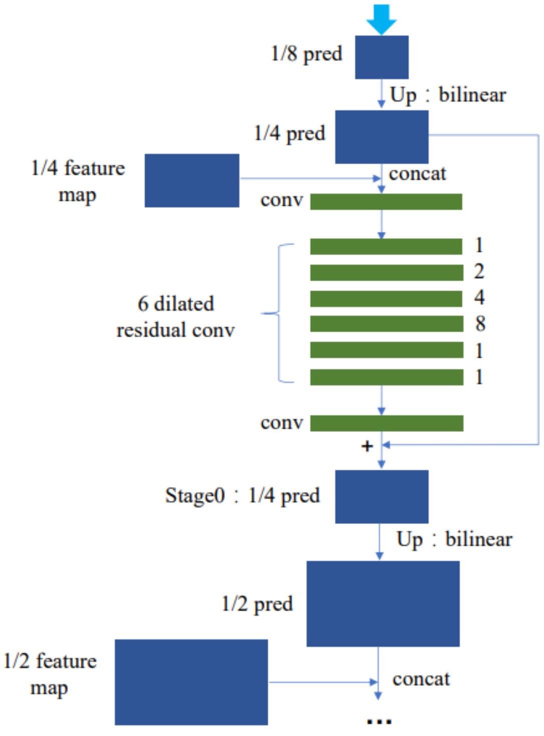
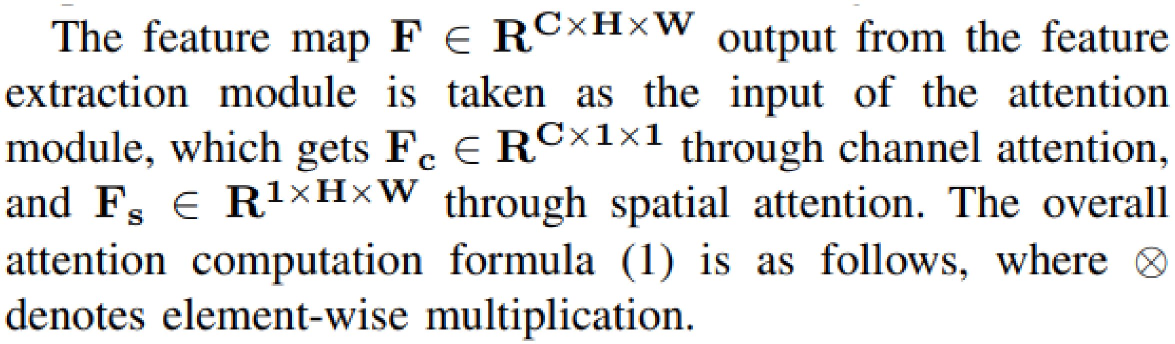
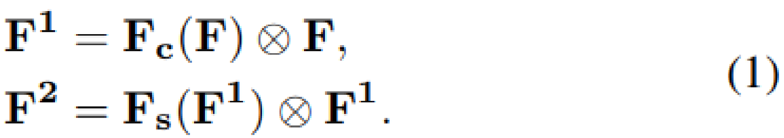


Fig. 3. Refinement structure overview of proposed ASMNet. We concatenate pre-stage features and coarse prediction, and then use 6 dialted convolution blocks with different dialed rates and output the optimized disparity map. Then the output of the upper level optimization is taken as the input of the next level optimization until the network outputs with full resolution as the final disparity prediction map.

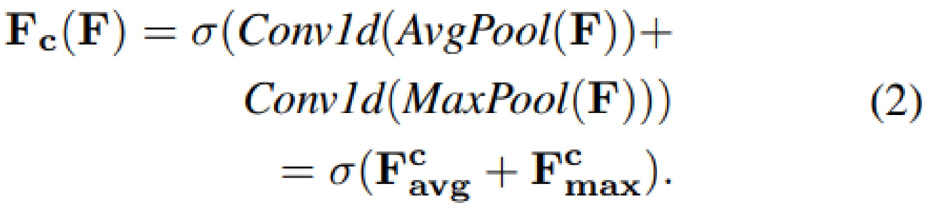
- **Our Method**



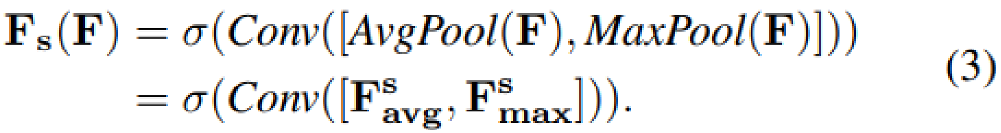
The feature map $\mathbf{F} \in \mathbf{R}^{\mathbf{C \times H \times W}}$ output from the feature extraction module is taken as the input of the attention module, which gets $\mathbf{F_c} \in \mathbf{R}^{\mathbf{C \times 1 \times 1}}$ through channel attention, and $\mathbf{F_s} \in \mathbf{R}^{\mathbf{1 \times H \times W}}$ through spatial attention. The overall attention computation formula (1) is as follows, where $\otimes$ denotes element-wise multiplication.



$$\begin{aligned} \mathbf{F^1} &= \mathbf{F_c}(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{F^2} &= \mathbf{F_s}(\mathbf{F^1}) \otimes \mathbf{F^1}. \end{aligned} \tag{1}$$



$$\begin{aligned} \mathbf{F_c}(\mathbf{F}) = \sigma(&Conv1d(AvgPool(\mathbf{F}))+ \\ &Conv1d(MaxPool(\mathbf{F}))) \\ = \sigma(&\mathbf{F^c_{avg}} + \mathbf{F^c_{max}}). \end{aligned} \tag{2}$$



$$\begin{aligned} \mathbf{F_s}(\mathbf{F}) &= \sigma(Conv([AvgPool(\mathbf{F}), MaxPool(\mathbf{F})])) \\ &= \sigma(Conv([\mathbf{F^s_{avg}}, \mathbf{F^s_{max}}])). \end{aligned} \tag{3}$$

- **Loss Function Selection: smoothL1**

$$L(d,\hat{d}) = \frac{1}{N}\sum_{i=1}^{N} smooth_{L_1}(d_i - \hat{d}_i), \qquad (5)$$

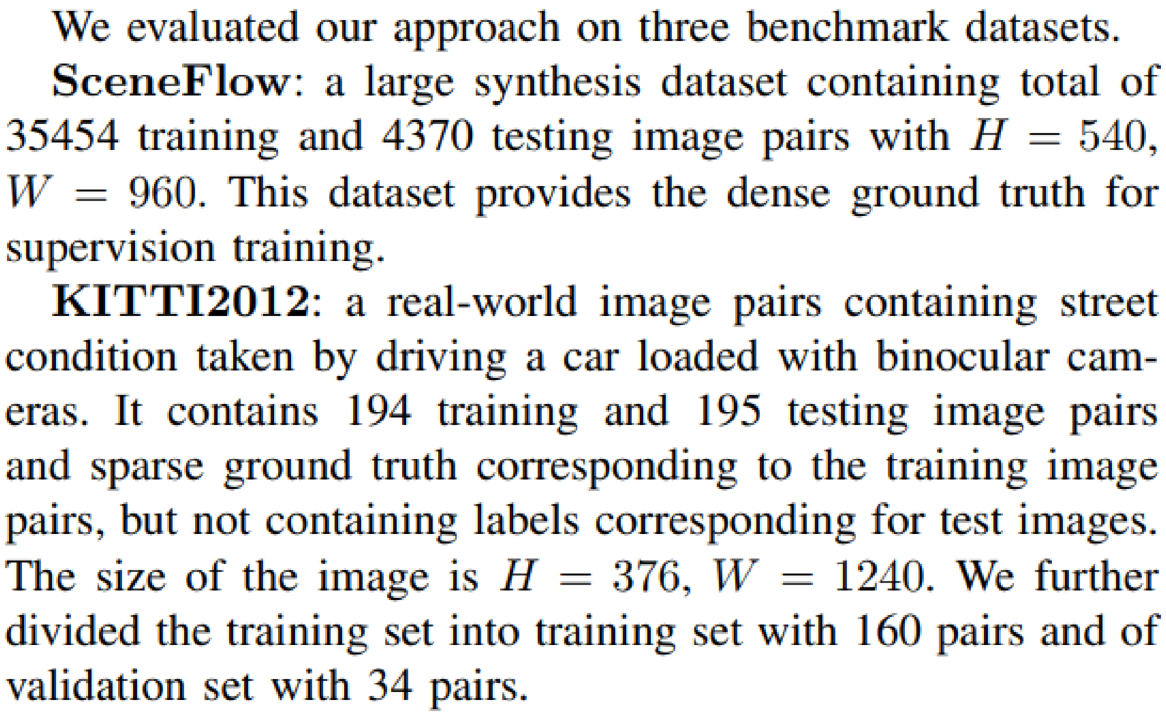
$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise }. \end{cases} \qquad (6)$$

$$L_2(x) = x^2 \qquad (1) \quad \longrightarrow \quad \frac{\mathrm{d}L_2(x)}{\mathrm{d}x} = 2x \qquad (4)$$

$$L_1(x) = |x| \qquad (2) \quad \longrightarrow \quad \frac{\mathrm{d}L_1(x)}{\mathrm{d}x} = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{otherwise} \end{cases} \qquad (5)$$

$$\mathrm{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \qquad (3) \quad \longrightarrow \quad \frac{\mathrm{d}\ \mathrm{smooth}_{L_1}}{\mathrm{d}x} = \begin{cases} x & \text{if } |x| < 1 \\ \pm 1 & \text{otherwise} \end{cases} \qquad (6)$$

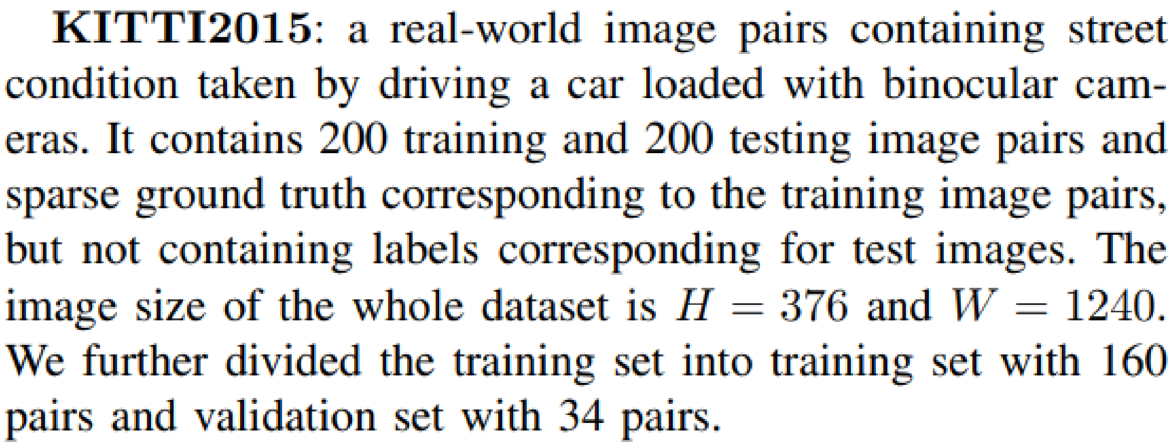
- **Datasets: SceneFlow KITTI2012 KITTI2015**



We evaluated our approach on three benchmark datasets.

SceneFlow: a large synthesis dataset containing total of 35454 training and 4370 testing image pairs with $H = 540$, $W = 960$. This dataset provides the dense ground truth for supervision training.

KITTI2012: a real-world image pairs containing street condition taken by driving a car loaded with binocular cameras. It contains 194 training and 195 testing image pairs and sparse ground truth corresponding to the training image pairs, but not containing labels corresponding for test images. The size of the image is $H = 376$, $W = 1240$. We further divided the training set into training set with 160 pairs and of validation set with 34 pairs.



KITTI2015: a real-world image pairs containing street condition taken by driving a car loaded with binocular cameras. It contains 200 training and 200 testing image pairs and sparse ground truth corresponding to the training image pairs, but not containing labels corresponding for test images. The image size of the whole dataset is $H = 376$ and $W = 1240$. We further divided the training set into training set with 160 pairs and validation set with 34 pairs.

- **Ablation Experiments**

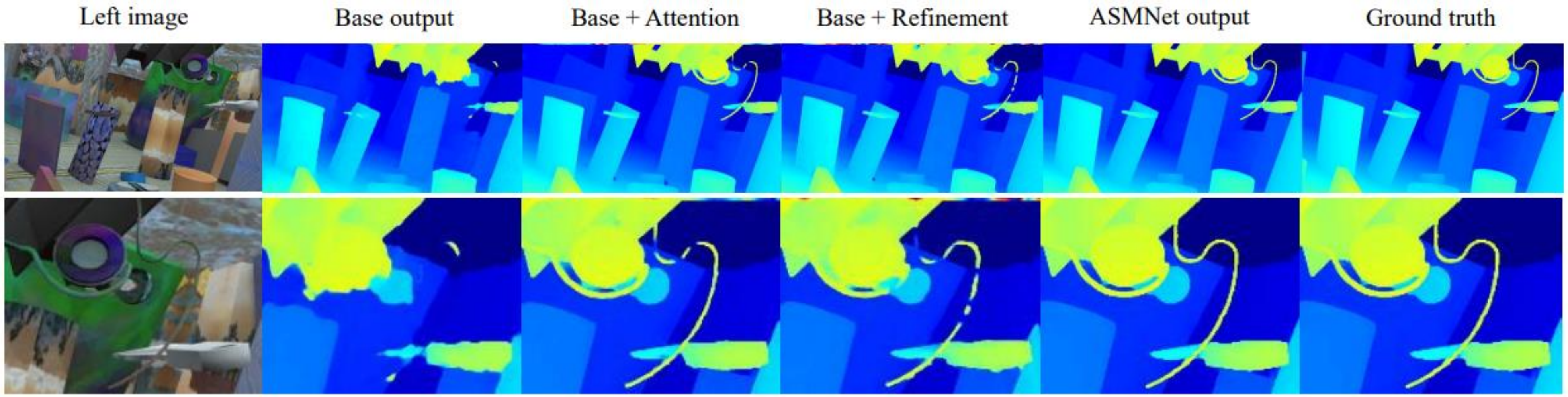


Fig. 4. Visualization of ablation experiments in slender structures.

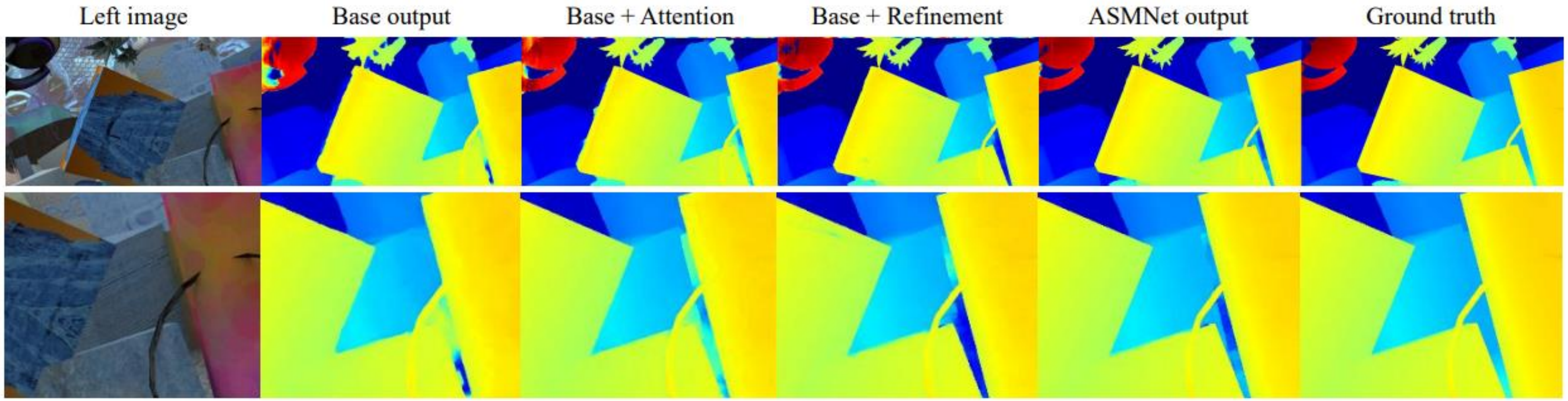
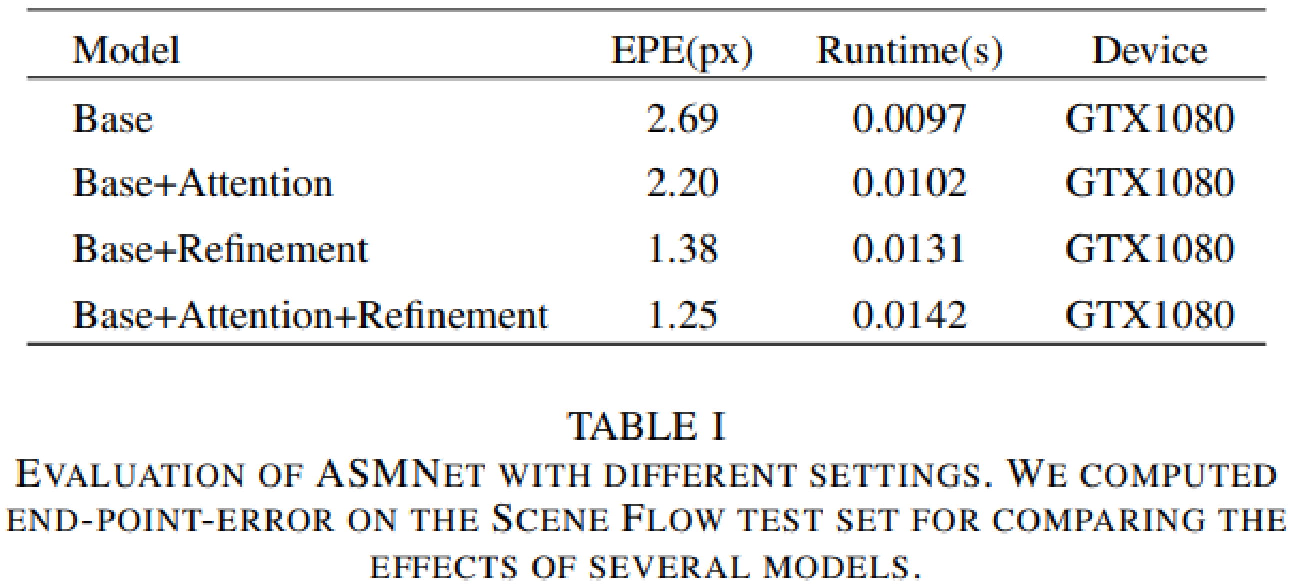


Fig. 5. Visualization of ablation experiments in textureless regions.

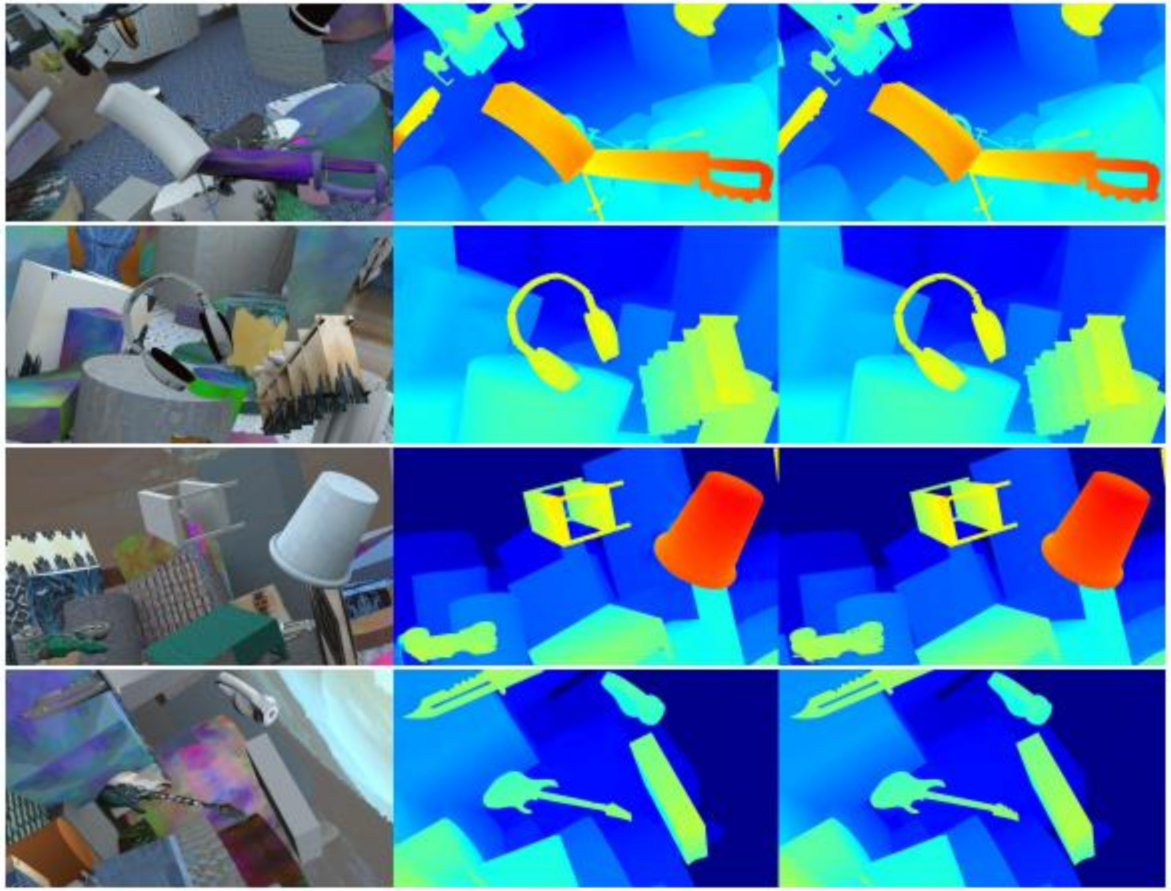
- **Ablation Experiments**



Model	EPE(px)	Runtime(s)	Device
Base	2.69	0.0097	GTX1080
Base+Attention	2.20	0.0102	GTX1080
Base+Refinement	1.38	0.0131	GTX1080
Base+Attention+Refinement	1.25	0.0142	GTX1080

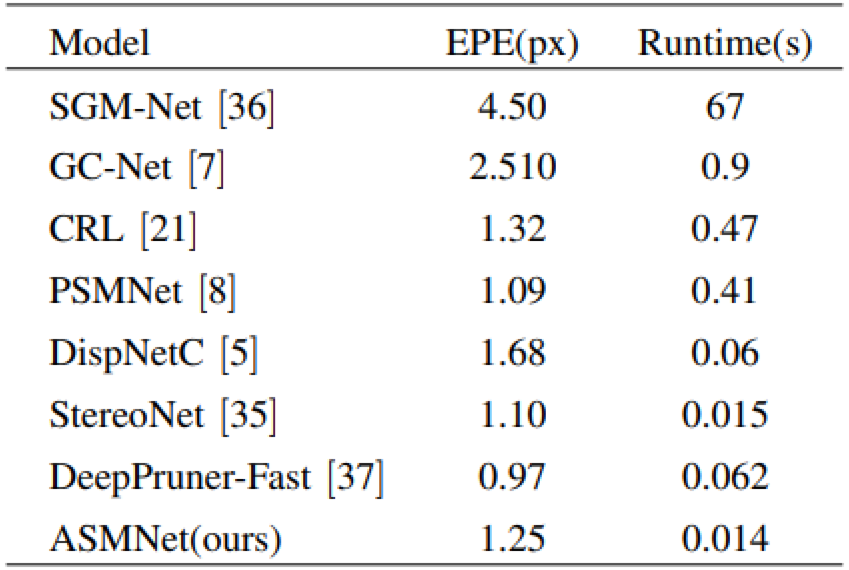
TABLE I
Evaluation of ASMNet with different settings. We computed end-point-error on the Scene Flow test set for comparing the effects of several models.

● Over All Experiments



Left image ASMNet output Ground truth

Fig. 6. Visualization results on the Scene Flow dataset.



Model	EPE(px)	Runtime(s)
SGM-Net [36]	4.50	67
GC-Net [7]	2.510	0.9
CRL [21]	1.32	0.47
PSMNet [8]	1.09	0.41
DispNetC [5]	1.68	0.06
StereoNet [35]	1.10	0.015
DeepPruner-Fast [37]	0.97	0.062
ASMNet(ours)	1.25	0.014

TABLE II
EVALUATION OF ASMNET WITH DIFFERENT SETTINGS. WE COMPUTED END-POINT-ERROR ON THE SCENE FLOW TEST SET, COMPARED THE EFFECTS OF SEVERAL MODELS.

- **Over All Experiments**

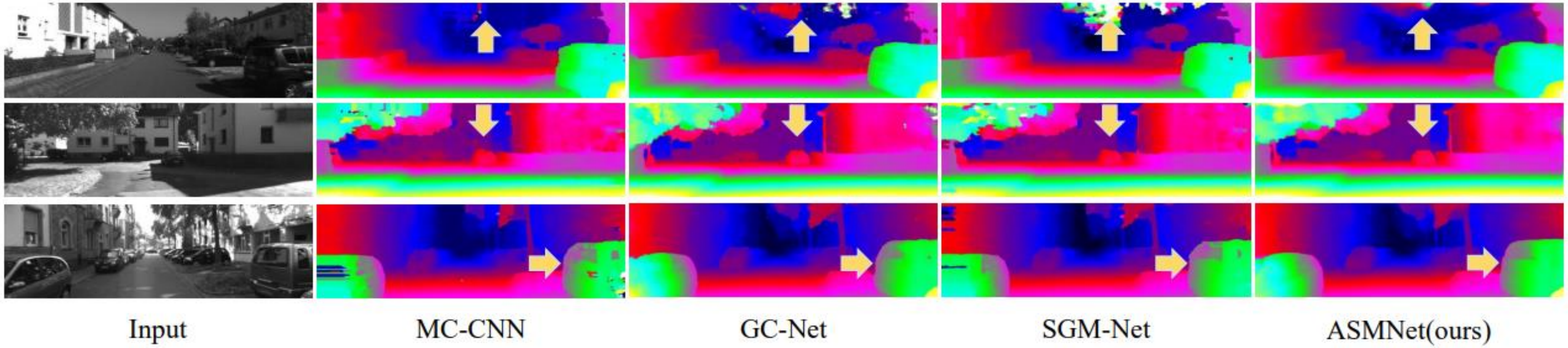


Fig. 7. Visualization results and comparisons on the KITTI2012 dataset. The left panel shows the left input image of stereo image pair. For each input image, the disparity maps obtained by MC-CNN [4], GC-Net [7], SGM-Net [36], and ASMNet are illustrated together.

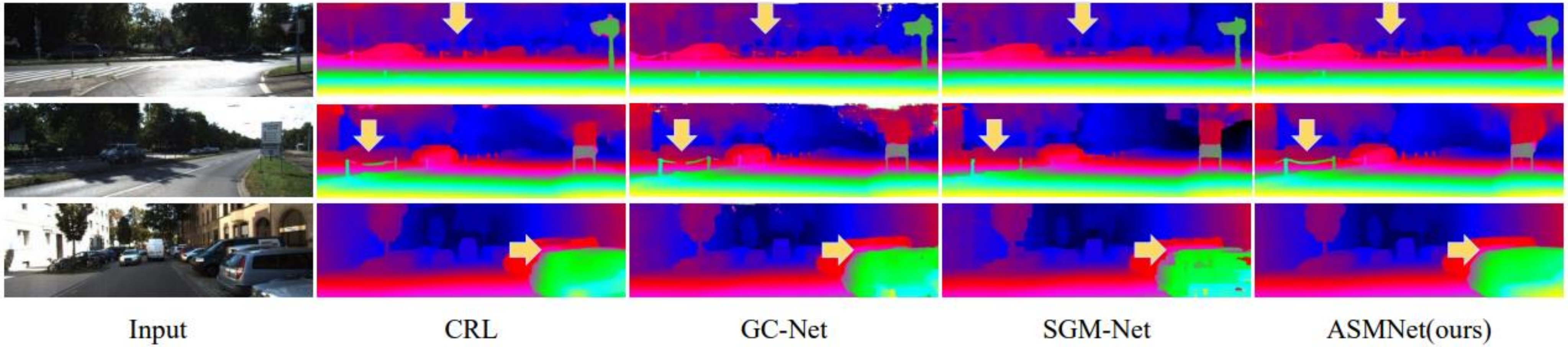
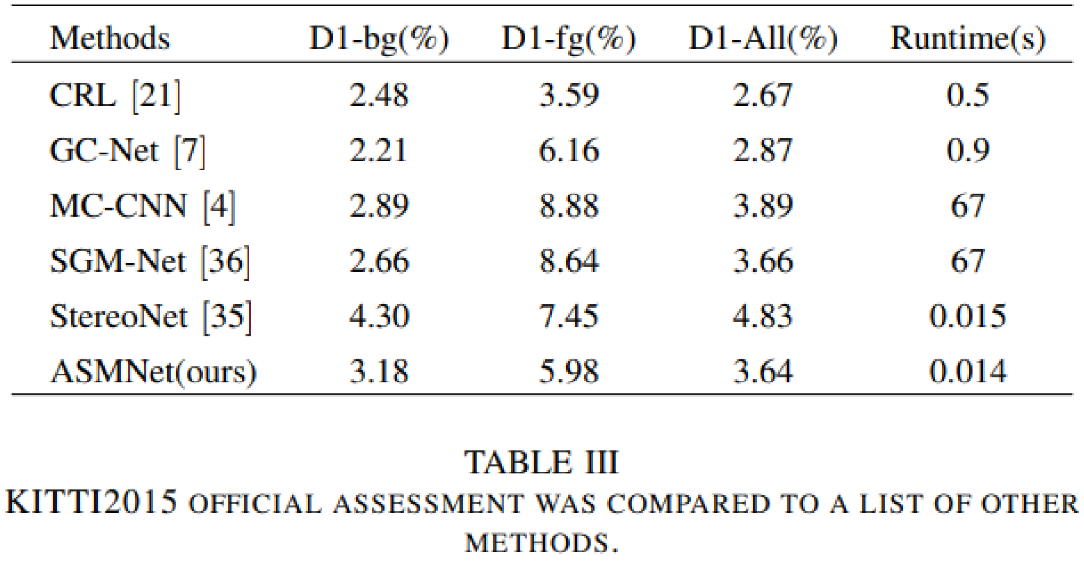


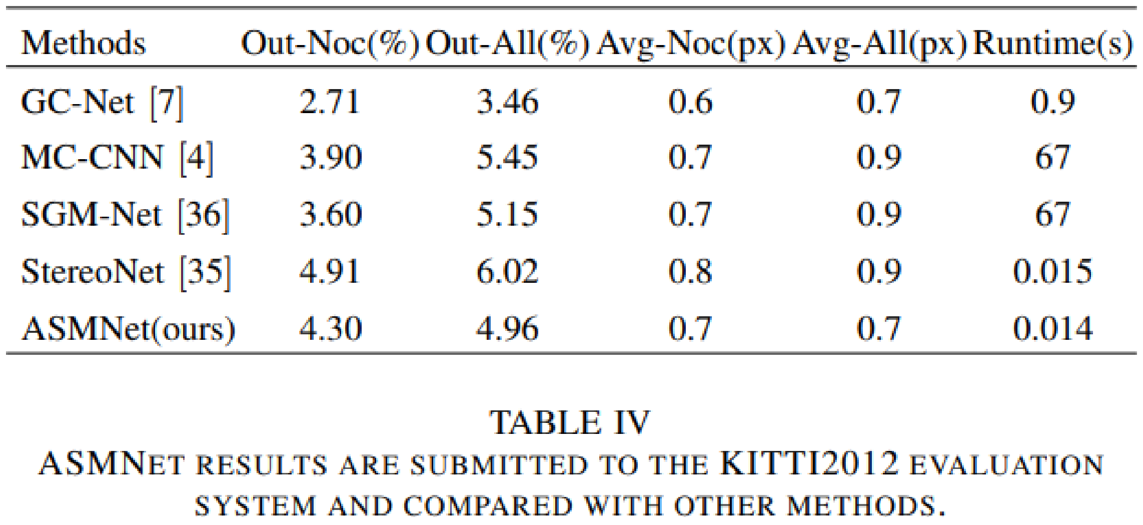
Fig. 8. Visualization results and comparisons on the KITTI2015 dataset. The left panel shows the left input image of stereo image pair. For each input image, the disparity maps obtained by CRL [21], GC-Net [7], SGM-Net [36], and ASMNet are illustrated together.

- **Over All Experiments**



Methods	D1-bg(%)	D1-fg(%)	D1-All(%)	Runtime(s)
CRL [21]	2.48	3.59	2.67	0.5
GC-Net [7]	2.21	6.16	2.87	0.9
MC-CNN [4]	2.89	8.88	3.89	67
SGM-Net [36]	2.66	8.64	3.66	67
StereoNet [35]	4.30	7.45	4.83	0.015
ASMNet(ours)	3.18	5.98	3.64	0.014

TABLE III
KITTI2015 OFFICIAL ASSESSMENT WAS COMPARED TO A LIST OF OTHER METHODS.



Methods	Out-Noc(%)	Out-All(%)	Avg-Noc(px)	Avg-All(px)	Runtime(s)
GC-Net [7]	2.71	3.46	0.6	0.7	0.9
MC-CNN [4]	3.90	5.45	0.7	0.9	67
SGM-Net [36]	3.60	5.15	0.7	0.9	67
StereoNet [35]	4.91	6.02	0.8	0.9	0.015
ASMNet(ours)	4.30	4.96	0.7	0.7	0.014

TABLE IV
ASMNET RESULTS ARE SUBMITTED TO THE KITTI2012 EVALUATION SYSTEM AND COMPARED WITH OTHER METHODS.

- **Conclusion**

In this work, we propose ASMNet, a novel end-to-end real time CNN architecture for stereo matching which consists of two main modules to exploit correlation information and disparity refinement: the attention module and the disparity refinement module. The attention module captures left and right feature maps correlation to form a cost volume. The disparity refinement module further learns to optimize the coarse prediction via introducing edge information and details by dilated convolutions with different dilated rates and enlargeing the coarse resolution ceaselessly until getting full-size disparity map. Experiments show the effectiveness of the proposed method.

Thank you for listening!