

Bilateral Grid Learning for Stereo Matching Networks

Bin Xu¹, Yuhua Xu^{1,2,*}, Xiaoli Yang¹, Wei Jia², Yulan Guo³

¹Orbbec, ²Hefei University of Technology, ³Sun Yat-sen University

xyh_nudt@163.com, guoyulan@sysu.edu.cn

Abstract

Real-time performance of stereo matching networks is important for many applications, such as automatic driving, robot navigation and augmented reality (AR). Although significant progress has been made in stereo matching networks in recent years, it is still challenging to balance real-time performance and accuracy. In this paper, we present a novel edge-preserving cost volume upsampling module based on the slicing operation in the learned bilateral grid. The slicing layer is parameter-free, which allows us to obtain a high quality cost volume of high resolution from a low-resolution cost volume under the guide of the learned guidance map efficiently. The proposed cost volume upsampling module can be seamlessly embedded into many existing stereo matching networks, such as GCNet, PSMNet, and GANet. The resulting networks are accelerated several times while maintaining comparable accuracy. Furthermore, we design a real-time network (named BGNet) based on this module, which outperforms existing published real-time deep stereo matching networks, as well as some complex networks on the KITTI stereo datasets. The code is available at <https://github.com/YuhuaXu/BGNet>.

1. Introduction

Stereo matching is a key step in 3D reconstruction, which has numerous applications in the fields of 3D modeling, robotics, UAVs, augmented realities (AR), and autonomous driving [27, 10, 1]. Given a pair of stereo images, the purpose of stereo matching is to establish dense correspondences between the pixels in the left and right images. Although this problem has been studied for more than 40 years, it has not been completely solved due to some difficult factors, such as sensor noise, foreground-background occlusion, weak or repeated textures, reflective regions, and transparent objects.

In recent years, deep learning has shown great potential in this field [43, 14, 18, 16, 5, 37]. Some recent work indi-

cates that 3D convolutions can improve the accuracy of the disparity estimation networks [14, 5, 9]. However, the 3D convolutions are time consuming, which limits their application in real-time systems.

Bilateral filter [32] is an edge-preserving filter, which has wide applications in image de-noising, disparity estimation [41], and depth upsampling [40]. However, the original implementation of bilateral filter is time-consuming. Bilateral grid proposed by Chen *et al.* [7] is a technique for speeding up bilateral filtering. It treats the filter as a “splat/blur/slice” procedure. That is, pixel values are “splatted” onto a small set of vertices in a grid, those vertex values are then blurred, and finally the filtered pixel values are produced via a “slice” (an interpolation) of the blurred vertex values. Recently, Gharbi *et al.* [11] utilize the idea of the bilateral grid to estimate the local affine color transforms in their real-time image enhancement network.

For StereoNet [16], the 2D disparity map regressed from an aggregated 4D cost volume at a low resolution (e.g., 1/8) is upsampled via bilinear interpolation and refined hierarchically. Although the speed of the lightweight network is fast, the accuracy is relatively low compared with existing complex networks, as shown in Table 4. On the ranking list of KITTI 2012 [10] and KITTI 2015 [23], the top performing methods usually conduct 3D convolutions at a relatively high resolution, such as 1/3 resolution for GANet [44] and 1/4 resolution for PSMNet [44]. However, their efficiency is reduced. For example, it takes 1.8s for GANet to process a pair of images of 1242×375 , and 0.41s for PSMNet.

The motivation of this paper is to propose a solution that can regress the disparity map at a high resolution to keep the high accuracy, while maintaining high efficiency. The main contributions of this work are as follows:

(1) We design a novel edge-preserving cost volume upsampling module based on the learned bilateral grid. It can efficiently obtain a high-resolution cost volume for disparity estimation from a low-resolution cost volume via a slicing operation. With this module, cost volume aggregations (e.g., 3D convolutions) can be performed at a low resolution. The proposed cost volume upsampling module can be seamlessly embedded into many existing networks such

*Corresponding author

as GCNet, PSMNet, and GANet. The resulting networks can be accelerated 4~29 times while maintaining comparable accuracy. To the best of our knowledge, it is the first time that the differential bilateral grid operation is applied in deep stereo matching networks.

(2) Based on the advantages of the proposed cost volume upsampling module, we design a real-time stereo matching network, named BGNet, which can process stereo pairs on the KITTI 2012 and KITTI 2015 datasets at 39fps. Experimental results show that BGNet outperforms existing published real-time deep stereo matching networks, as well as some complex networks, such as GCNet, AANet, DeepPruner-Fast, and FADNet, on the KITTI 2012 and KITTI 2015 stereo datasets.

2. Related Work

Stereo Matching Network. MC-CNN [42] is the first work that uses convolutional neural network (CNN) to compare two image patches (e.g., 11×11) and calculate their matching costs. Meanwhile, the following steps, such as cost aggregation, disparity computation, and disparity refinement, are still traditional methods [22]. It significantly improves the accuracy, but still struggles to produce accurate disparity results in textureless, reflective and occluded regions and is time-consuming. DispNetC [20] is the first end-to-end stereo matching network with a similar network structure as FlowNet [8]. DispNetC is more efficient, almost 1000 times faster than MC-CNN-Acrt [42]. In DispNetC, there is an explicit correlation layer. To further improve the estimation accuracy, the residual refinement layers are exploited [18, 19, 24]. Besides, the segmentation information [39] and the edge information [31] are incorporated into the networks to improve the performance. GCNet [15] uses 3D convolutions for cost aggregation in a 4D cost volume, and utilizes the soft *argmin* to regress the disparity. DeepPruner [9] brings the idea of PatchMatch Stereo [3], and builds a narrow cost volume based on the estimated lower and upper bounds of the disparity to speed up the prediction. The narrow cost volume optimization is also used in [12]. Since disparities can vary significantly for stereo cameras with different baselines, focal lengths and resolutions, the fixed maximum disparity used in cost volume hinders them to handle different stereo image pairs with large disparity variations. Wang *et al.* [34, 33] propose a generic parallax-attention mechanism to capture stereo correspondence regardless of disparity variations.

Recent work [5, 9] shows that the networks with 3D convolution can achieve higher disparity estimation accuracy on specific datasets. However, 3D convolution is more time-consuming than 2D convolution, which makes it difficult to apply in real-time applications. In order to pursue real-time performance, StereoNet [16] performs 3D convolution at a low resolution (e.g., 1/8 resolution) and the resulting

network can run in real-time at 60fps. However, this simplification reduces the network’s accuracy. For other networks, such as AANet [37], FADNet [35], and DeepPruner-Fast [9], although the accuracy has been improved, they have not achieved real-time performance.

Bilateral Grid. Since this work is inspired by bilateral grid of Chen *et al.* [7], we will give it a brief review.

The bilateral grid was originally introduced to speed up the bilateral filter. It consists of three steps, including splatting, blurring, and slicing. For an input image I and a guidance image G , the splatting operation projects the original pixels of I into a 3D grid \mathcal{B} , where the first two dimensions (x, y) correspond to 2D position in the image plane and form the spatial domain, while the third dimension g corresponds to the image intensity of the guidance image. Then, Gaussian blurring is performed in the 3D bilateral grid \mathcal{B} . Finally, based on the blurred bilateral grid and the guidance image G , a 2D value map \bar{I} is extracted by accessing the grid at $(sx, sy, s_G G(x, y))$ using trilinear interpolation, where s is the width or height ratio of the grid’s dimension w.r.t the original image dimension, and $s_G \in (0, 1)$ is the ratio of the gray level of the grid to the gray level of the guidance image G . This linear interpolation under the guide of the guidance image in the bilateral grid is called slicing. In practice, most operations on the grid require only a coarse resolution, where the number of grid cells is much smaller than the number of image pixels. See [7] for more details.

The bilateral grid is utilized to accelerate stereo matching algorithms [26, 2]. Chen *et al.* [6] approximate an image operator with a grid of local affine models in bilateral space, the parameters of which are fit to a reference input and output pair. By performing this model-fitting on a low-resolution image pair, this technique enables real-time on-device computation. Gharbi *et al.* [11] build upon this bilateral space representation. Rather than fitting a model to approximate a single instance of an operator from a pair of images, they construct a rich CNN-like model that is trained to apply the operator to any unseen input.

3. Method

Inspired by bilateral grid processing [7], we propose an edge-aware cost volume upsampling module. With this module, we can perform the majority of calculation at a low resolution. Meanwhile, we can obtain accurate disparity prediction with the upsampled cost volume at a high resolution. In this section, we first describe the proposed cost volume upsampling module in details. Then, we show that the upsampling module can be utilized as an embedded module in many existing stereo matching networks. Additionally, based on the advantages of this module, we design a real-time stereo matching network.

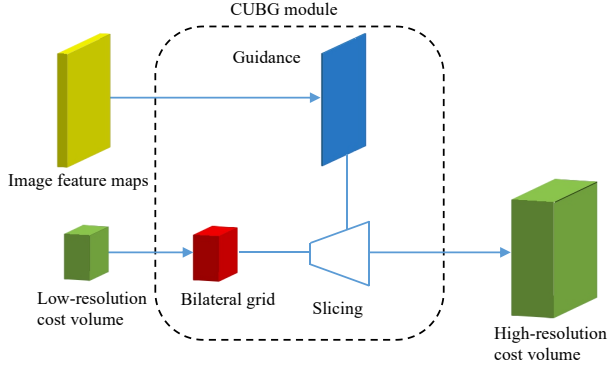


Figure 1. The module of Cost volume Upsampling in the learned Bilateral Grid (CUBG). The high-quality cost volume at high resolution can be obtained with the CUBG module from the low resolution (e.g., 1/8) via the slicing operation of bilateral grid processing.

3.1. Cost Volume Upsampling in Bilateral Grid

As stated in Section 1, the goal of this paper is to seek a solution that can regress the disparity map at a high resolution to keep the high accuracy, while maintaining high efficiency. In order to achieve this goal, we design a module of Cost volume Upsampling in Bilateral Grid (CUBG). The CUBG module can upsample the cost volume calculated at the low resolution (e.g., 1/8) to a high resolution via the slicing operation of bilateral grid processing.

As illustrated in Figure 1, the input of the CUBG module is a low-resolution cost volume \mathcal{C}_L and image feature maps, and the output is the upsampled high-resolution cost volume. The operation in the CUBG module includes bilateral grid creation and slicing.

Cost volume as a bilateral grid. Given an aggregated cost volume \mathcal{C}_L with four dimensions (including width x , height y , disparity d , and channel c) at a low resolution (e.g., 1/8), the conversion from the cost volume \mathcal{C}_L to the bilateral grid \mathcal{B} is straightforward. In all experiments, we use a 3D convolution of $3 \times 3 \times 3$ to achieve the conversion.

There are four dimensions in the bilateral grid \mathcal{B} : width x , height y , disparity d , and guidance feature g . The value in the bilateral grid is represented by $\mathcal{B}(x, y, d, g)$.

Upsampling with a slicing layer. With the bilateral grid, we can produce a 3D high-resolution cost volume \mathcal{C}_H ($\mathcal{C}_H \in \mathbb{R}^{W,H,D}$) via a slicing layer. This layer performs data-dependent lookups in the low-resolution grid of matching costs. Specifically, the slicing operation is the linear interpolation in the four-dimensional bilateral grid under the guide of the 2D guidance map G of high resolution. The slicing layer is parameter-free and can be implemented efficiently. Formally, the slicing operation is defined as

$$\mathcal{C}_H(x, y, d) = \mathcal{B}(sx, sy, sd, s_G G(x, y)) \quad (1)$$

where $s \in (0, 1)$ is the width or height ratio of the grid's dimension w.r.t the high-resolution cost volume dimension, $s_G \in (0, 1)$ is the ratio of the gray level of the grid (l_{grid}) to the gray level of the guidance map l_{guid} .

The guidance map G is generated from the high-resolution feature maps via two 1×1 convolutions. Note that, the guidance information of each pixel depends on its own feature vector. Consequently, sharp edges can be obtained.

Unlike the original grid designed in [7], the bilateral grid in this work is learned from the cost volume automatically. In experiments, the size of the grid is usually set to $W/8 \times H/8 \times D_{max}/8 \times 32$, where W and H are image width and height, respectively. D_{max} is the maximal disparity value.

3.2. Network Architecture

3.2.1 Embedded Module

The proposed CUBG module can be seamlessly embedded into many existing end-to-end network architectures. In this work, we embed the CUBG module into four representative stereo matching models: GCNet [14], PSMNet [5], GANet [44], and DeepPrunerFast [9]. The resulting models are denoted with a suffix BG (e.g., GCNet-BG). For the first three networks, we re-build the cost volume at the resolution of 1/8, 1/8, and 1/6, respectively. Then, the cost volumes are upsampled via our CUBG module to the resolution of 1/2, 1/4, and 1/3, respectively. For DeepPrunerFast, the PatchMatch-like upper and lower bounds estimation module and the narrow cost volume aggregation module are replaced by a full cost volume aggregation at the resolution of 1/8. Then, the cost volume is upsampled via the CUBG module to the resolution of 1/2. All the other parts remain the same as in their original implementation.

3.2.2 BGNet

Based on the CUBG module, we also design an efficient end-to-end stereo matching network, named BGNet. Figure 2 provides an overview of the network. The network consists of four modules of feature extraction, cost volume aggregation, cost volume upsampling, and residual disparity refinement. It can run in real-time at the resolution of the KITTI stereo dataset. In the following, we introduce these modules in details.

Feature Extraction. The ResNet-like architecture is widely used in stereo matching networks [5, 13]. We utilize the similar architecture to extract the image features for matching here. For the first three layers, three convolution of 3×3 kernel with strides of 2, 1, and 1 are used to downsample the input images. Then, four residual layers with strides of 1, 2, 2, and 1 are followed to quickly produce unary features at 1/8 resolution. Two hourglass networks are followed to obtain large receptive fields and rich

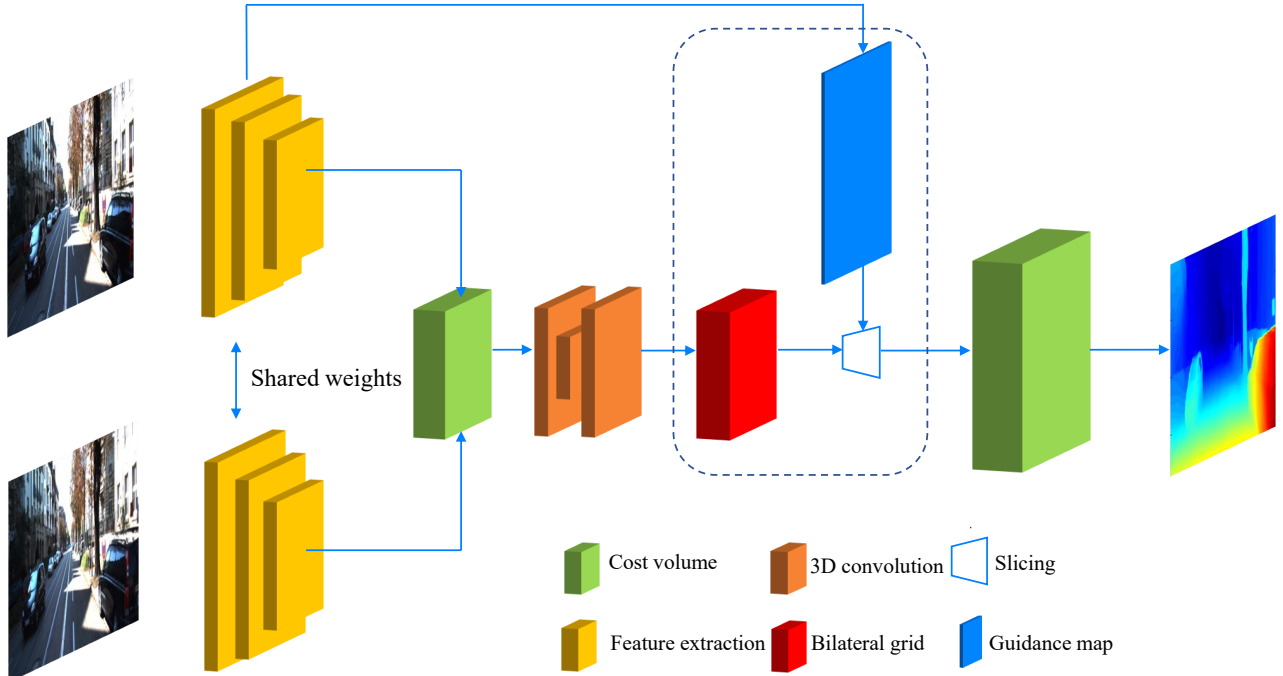


Figure 2. Overview of the proposed BGNet.

semantic information. Finally, all the feature maps at 1/8 resolution are concatenated to form feature maps with 352 channels for the generation of the cost volume. We use \mathbf{f}_l and \mathbf{f}_r to represent the final feature maps extracted from the left and right images, respectively.

Cost Aggregation. After the feature extraction modules, we build a group-wise correlation cost volume [13] for aggregation, which combines the advantages of the concatenation volume and the correlation volume. Group-wise correlation is computed for each pixel location (x, y) at disparity level d by dividing the feature channels into $N_{\bar{g}}$ groups, the same as in [13]. In our experiments, $N_{\bar{g}} = 44$.

In PSMNet [5], a stacked hourglass architecture was utilized to optimize the cost volume using the contextual information. Here, considering the efficiency, only one hourglass architecture is used to filter the cost volume. Specifically, we first use two 3D convolutions to reduce the channel number of cost volume from 44 to 16. Then, a U-Net-like 3D convolution network is used for cost aggregation, where skip connection is replaced by an element-wise summation operation to reduce computational cost. We use $\mathcal{C}_{\mathcal{L}}$ to represent the aggregated cost volume at 1/8 resolution.

Disparity Regression. With the high-resolution cost volume \mathcal{C}_H , we can regress the disparity prediction via *softmax* as in [14]:

$$\mathbf{D}_{pred}(x, y) = \sum_{d=0}^{D_{max}} d \times \text{softmax}(\mathcal{C}_H(x, y, d)) \quad (2)$$

Loss Function. The loss function L is defined on the final prediction \mathbf{D}_{pred} using the smooth L_1 loss \mathcal{L} ,

$$L = \sum_p \mathcal{L}(\mathbf{D}_{pred}(p) - \mathbf{D}_{gt}(p)) \quad (3)$$

Here,

$$\mathcal{L}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases}$$

where $\mathbf{D}_{gt}(p)$ is the ground truth disparity for pixel p .

4. Experimental Results

4.1. Datasets and Evaluation Metrics

SceneFlow. The synthetic stereo datasets include Flyingthings3D, Driving, and Monkaa. These datasets consist of 35,454 training images and 4,370 testing images of size 960×540 with accurate ground-truth disparity maps. Since the Finalpass of the SceneFlow datasets contains more motion blur and defocus and is more like real-world images than the Cleanpass, we use the Finalpass for ablation study. The End-Point-Error (EPE) will be used as the evaluation metric for the SceneFlow dataset.

KITTI. KITTI 2012 [10] and KITTI 2015 [23] are outdoor driving scene datasets. KITTI 2012 provides 194 training and 195 testing image pairs, KITTI 2015 provides 200 training and 200 testing image pairs. The resolution of

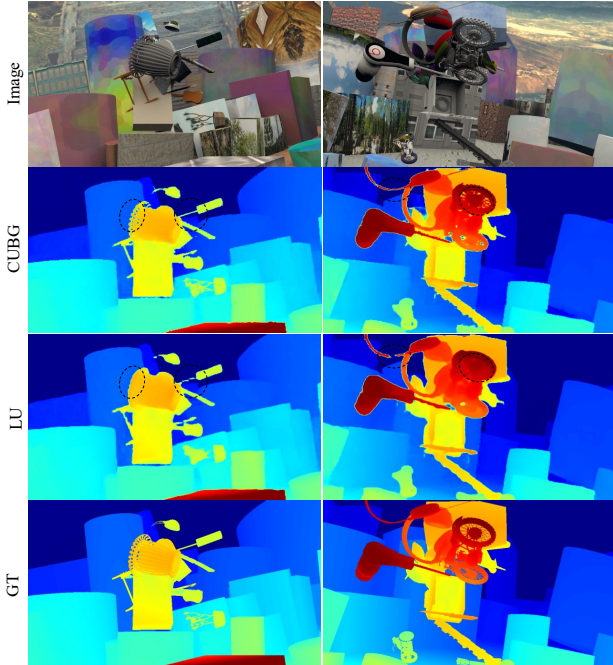


Figure 3. Qualitative results on SceneFlow. The abbreviations Linear Upsampling (LU) and CUBG denote the two different cost volume upsampling methods.

KITTI 2015 is 1242×375 , and that of KITTI 2012 is 1226×370 . The groundtruth disparities are obtained from LiDAR points. For KITTI 2012, percentages of erroneous pixels and average end-point errors for both nonoccluded (*Noc*) and all (*All*) pixels are reported. For KITTI 2015, the percentage of disparity outliers *DI* is evaluated.

Middlebury 2014. The Middlebury stereo dataset [28] had been widely used for the evaluation of stereo matching methods. The disparity maps in this dataset are calculated using accurate structured-light techniques. However, this dataset only contains dozens of image pairs and is insufficient to train a deep neural network.

We follow previous end-to-end approaches by performing initial training on SceneFlow and then individually finetuning the resulting model on the KITTI datasets. For Middlebury, we only use the dataset for the evaluation of generalization ability without fine-tuning.

4.2. Implementation Details

Our BGNet is implemented with PyTorch on NVIDIA RTX 2080Ti GPU. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

For the SceneFlow dataset, a series of data augmentation techniques, including asymmetric chromatic augmentation, y-disparity augmentation [38], Gaussian blur enhancement, and scale zoom enhancement are adopted. Each enhance-

ment module has a 50% chance of being used. In addition, following CRL [25], the stereo pairs with more than 25% of their disparity values larger than 300 are removed. We use the one-cycle scheduler [30] to adjust the learning rate with the maximum value of 0.001. The batch size is set to 16, and crop size is set to 512×256 to train our network for 50 epochs in total. To evaluate these methods on the test dataset of SceneFlow, we set the maximum disparity values to 192. Pixels with disparity values out of the valid range are not considered in the evaluation.

For KITTI 2015, we randomly select 20 pairs as the validation set, and use the remaining 180 pairs and 194 pairs of KITTI 2012 as the training set. We first finetune our network with a constant learning rate of 0.001 for 300 epochs on the model pre-trained on SceneFlow, and repeat three times to pick the one with the best evaluation metrics. For KITTI 2012, we use a similar finetuning strategy.

4.3. Ablation Study

To validate the effectiveness of the proposed CUBG module, we first replace the module in BGNet with direct linear upsampling on the cost volume, and evaluate the EPE results on SceneFlow. As shown in Table 1, EPE rises from **1.17** to **1.40**. Qualitative results on SceneFlow are shown in Figure 3. Although the disparity map is reconstructed from a low-resolution cost volume, many thin structures and sharp edges are still recovered by BGNet, which benefits from the edge-aware nature of the bilateral grid. Performing disparity regression in a high-resolution cost volume obtained via a slicing layer in the bilateral grid forces the predictions of our network to follow the edges in the guidance map *G*, thereby regularizing our predictions towards edge-aware solutions.

To validate this point, we also evaluate the methods in flat regions and the regions near edges. Edges are first detected with the Canny detector [4] in ground-truth disparity maps. Then, these detected edges are dilated with a 5×5 square structuring element. The EPE of the regions of these dilated edges is denoted as EPE-edge, others are denoted as EPE-flat. Quantitative results are shown in Table 1, from which we have two major observations. First, for both methods, EPE-edge is much higher than EPE-flat, which means the disparities near edges are hard to estimate. Second, in flat regions, the errors of these two methods are comparable. However, in the regions near edges, the EPE of CUBG is 2.18 lower than linear upsampling, which benefits from the edge-aware property of CUBG. Figure 3 shows qualitative comparison. When the CUBG module is utilized, the predicted edges are more accurate and sharper than linear upsampling, and the thin structures are better reconstructed.

We also performed ablation study on Middlebury 2014 and KITTI 2015. For Middlebury 2014, we use 13 addi-

Method	EPE	EPE-edge	EPE-flat	Time (ms)
CUBG	1.17	5.95	0.68	25.3
LU	1.40	8.13	0.71	25.1

Table 1. Ablation study results of the proposed networks on Final-pass of the SceneFlow datasets [20]. The abbreviation LU denotes linear upsampling. EPE-edge represents the EPE near edges of objects, and EPE-flat represents the EPE in flat regions.

Method	Res-CV	EPE	Time (ms)
GCNet [14]	1/2	2.51	1673.2
GCNet-BG	1/8	1.07	57.1
PSMNet [5]	1/4	1.09	439.6
PSMNet-BG	1/8	0.92	89.4
GANet_deep [44]	1/3	0.95	2240.2
GANet_deep-BG	1/6	0.63	533.4
DeepPrunerFast [9]	1/8	0.97	64.7
DeepPrunerFast-BG	1/8	0.84	56.6

Table 2. Evaluation of the CUBG module embedded into other networks on the SceneFlow dataset [20]. We embed the CUBG module into four representative stereo matching models, GCNet, PSMNet, GANet, and DeepPrunerFast. Resulting models are denoted with a suffix BG. Res-CV represents the resolution at which the cost volume is built.

tional datasets with GT for training (finetuning), where 78 pairs of images are available in total. For the KITTI 2015 dataset, where the training set is split into 160 pairs for training and 40 pairs for validation (as done in AANet [37]). For Middlebury 2014, the Bad 2.0 errors for LU and BG are 18.7% and 16.8% (48.9% vs 45.3% near edges, and 14.7% vs 13.0% in flat regions). For KITTI 2015, the D1-all errors for LU and BG are 2.14% and 2.01%, respectively.

Impact of guidance map. If the original guidance map is replaced with the luma-version of the input image, the EPE on Scene Flow increases from 1.17 to 1.28.

Building cost volume at different resolutions. When building the cost volume at the resolution of 1/16, the EPE on Scene Flow increases from 1.17 to 1.58, and the run-time decreases from 25.3 ms to 17.1 ms.

Embedded module. Table 2 shows the quantitative results of GCNet-BG, PSMNet-BG, GANet-BG, and DeepPruner-BG, where the CUBG module is used as an embedded module in these networks. Compared with top-performing stereo models GC-Net, PSMNet and GANet, our method not only obtains clear performance improvements, but is also significantly faster (29× than GC-Net, 4.9× than PSMNet, and 4.2× than GANet), demonstrating the high efficiency of the CUBG module. Additionally, although the cost volumes of DeepPrunerFast and DeepPrunerFast-BG have the same resolution, DeepPrunerFast-BG is faster while having lower EPE.

Method	Res	D1-all (KIT)	T (ms) (KIT)	Bad2.0 (MID)
PSMNet [5]	1/4	1.95	410	19.36
PSMNet-BG	1/8	2.07	79	20.77
GANet_deep [44]	1/3	1.58	1800	15.67
GANet_deep-BG	1/6	1.67	406	15.30
DeepPrunerFast [9]	1/8	2.06	61	16.80
DeepPrunerFast-BG	1/8	1.91	55	15.14

Table 3. Evaluation of the CUBG module when embedded into three networks on the KITTI 2015 and Middlebury 2014 datasets. The second column gives the resolution at which the cost volume is built, the third and fourth columns show D1-all and run-time results on KITTI 2015 (KIT), and the last column shows Bad 2.0 errors on Middlebury (MID).

We also tested them on the KITTI 2015 and Middlebury 2014 datasets, where the finetuning strategies are the same as those in Subsection 4.3. The results in Table 3 show that, compared with their original networks, GA-Net-BG and PSM-Net-BG achieve comparable accuracy while being accelerated more than four times.

4.4. Evaluation on KITTI

Table 4 shows the performance and runtime of competing algorithms on the KITTI stereo benchmark. Our BGNet can run in real-time on the dataset at **39** fps. Compared with the latest FADNet [35], AANet [37] and DeepPrunerFast [9], our BGNet is not only more accurate, but also two times faster than those methods, as shown in Table 4. Among the published networks with computational time less than 50 ms, our BGNet has the best accuracy.

In addition, we build another model variant BGNet+. Compared with BGNet, BGNet+ has an additional hour-glass disparity refinement module as in [37]. The resulting model is still real-time (**30** fps). Although the model is light-weight, it even outperforms some complex networks, such as GCNet [14], iResNet [18] and PSMNet [5] on the KITTI 2015 dataset. Qualitative results of BGNet+ on KITTI 2015 are shown in Figure 4.

Runtime analysis. We calculated the average time of each module by testing 100 pairs of stereo images on KITTI 2015, as shown in Table 5. The proposed cost volume upsampling module is efficient and takes 4.3 ms.

The run-time for bilateral grid generation, guidance image calculation and upsampling with BG is 4.3 ms, 0.07 ms, and 0.74 ms, respectively. The linear cost volume upsampling (LU) takes 3.78 ms. However, when the whole network is deployed in GPU, BG upsampling is only 0.2 ms slower than linear upsampling. Note, the run-time of the whole network is not simply the sum of run-times consumed by all modules, this may be related to the parallel computing mechanism of GPU.

Method	KITTI 2012 [10]						KITTI 2015 [23]			Runtime (ms)
	2-noc	2-all	3-noc	3-all	EPE noc	EPE all	D1-bg	D1-fg	D1-all	
MC-CNN-acrt [43]	3.90	5.45	2.09	3.22	0.6	0.7	2.89	8.88	3.89	67000
SGM-Net [29]	3.60	5.15	2.29	3.50	0.7	0.9	2.66	8.64	3.66	67000
GANet [44]	1.89	2.50	1.19	1.6	0.4	0.5	1.48	3.46	1.81	1800
PSMNet [5]	2.44	3.01	1.49	1.89	0.5	0.6	1.86	4.62	2.32	410
GC-Net [14]	2.71	3.46	1.77	2.30	0.6	0.7	2.21	6.16	2.87	900
EdgeStereo-V2 [31]	2.32	2.88	1.46	1.83	0.4	0.5	1.84	3.30	2.08	320
iResNet2-i2 [18]	2.69	3.34	1.71	2.16	0.5	0.6	2.25	3.40	2.44	120
DeepPruner-Best[9]	-	-	-	-	-	-	1.87	3.56	2.15	182
DeepPruner-Fast[9]	-	-	-	-	-	-	2.32	3.91	2.59	61
RTSNet [17]	3.98	4.61	2.43	2.90	0.7	0.7	2.86	6.19	3.41	20
StereoNet [16]	4.91	6.02	-	-	0.8	0.9	4.30	7.45	4.83	15
DispNet [21]	7.38	8.11	4.11	4.65	0.9	1.0	4.32	4.41	4.34	60
FADNet[35]	3.98	4.63	2.42	2.86	0.6	0.7	2.68	3.50	2.82	50
AANet[37]	2.90	3.60	1.91	2.42	0.5	0.6	1.99	5.39	2.55	62
BGNet	3.13	3.69	1.77	2.15	0.6	0.6	2.07	4.74	2.51	25.4
BGNet+	2.78	3.35	1.62	2.03	0.5	0.6	1.81	4.09	2.19	32.3

Table 4. Quantitative evaluation on the test sets of KITTI 2012 and KITTI 2015. For KITTI 2012, we report the percentage of pixels with errors larger than x disparities in both non-occluded (x-noc) and all regions (x-all), as well as the overall EPE in both non occluded (EPE-noc) and all the pixels (EPE-all). For KITTI 2015, we report the percentage of pixels with EPEs larger than 3 pixels in background regions (D1-bg), foreground regions (D1-fg), and all (D1-all).

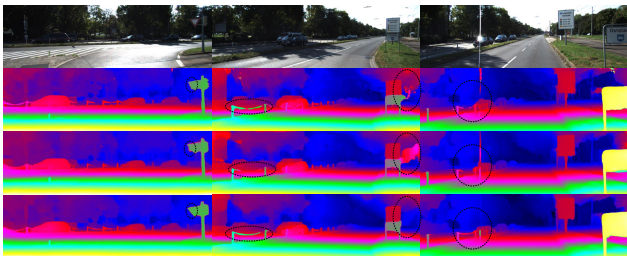


Figure 4. Qualitative comparisons on KITTI 2015. The first row shows the input RGB images, the second row shows the results of PSMNet, the third shows the results of DeepPruner-Fast, and the last row is the results of BGNet+.

Module	Time (ms)
Feature extraction	8.8
Cost volume building and aggregation	12.2
Bilateral grid operation	4.3
Refinement (for BGNet+)	7.0
Total	32.3

Table 5. Runtime analysis for each module of BGNet and BGNet+ on the KITTI 2015 dataset.

4.5. Generalization Performance

The generalization performance is important for a stereo network. We evaluate the generalization ability of our network on the training set of the Middlebury 2014 dataset, where the parameters of these networks are trained on Fly-

ingthings3D of SceneFlow [20] only, no additional training is done on Middlebury. The results are illustrated in Table 6. Our BGNet and BGNet+ show better performance than some complex networks, such as iResNet [18], GANet [44], and PSMNet [5]. Figure 5 shows the qualitative disparity estimation results achieved by BGNet+ and PSMNet on this benchmark.

For DeepPruner-Fast-BG, as shown in Table 6, the Bad 2.0 error is lower than its original version, which further demonstrates the effectiveness of our cost volume upsampling module.

DSMNet [45] has the best generalization performance, which benefits from feature normalization along both the spatial axis and the channel dimension. However, the domain invariant normalization process increases the computational time of the network. For the resolution of KITTI 2015, when the batch normalization modules are replaced by the domain invariant normalization modules, the computational time of the network is increased by 15.4 ms.

Additionally, we conduct another interesting evaluation. We use the IRS dataset [36], a large synthetic stereo dataset, and Flyingthings3D together to train BGNet and BGNet+. IRS contains more than 100,000 pairs of 960×540 resolution stereo images (84,946 for training and 15,079 for testing) in indoor scenes. The Bad 2.0 error of BGNet on Middlebury reduces from **17.5** to **13.5**, and the Bad 2.0 error of BGNet+ reduces from **17.2** to **11.6**. This shows that the synthetic dataset also plays a significant role in improving the generalization performance of the network without increasing computational time.

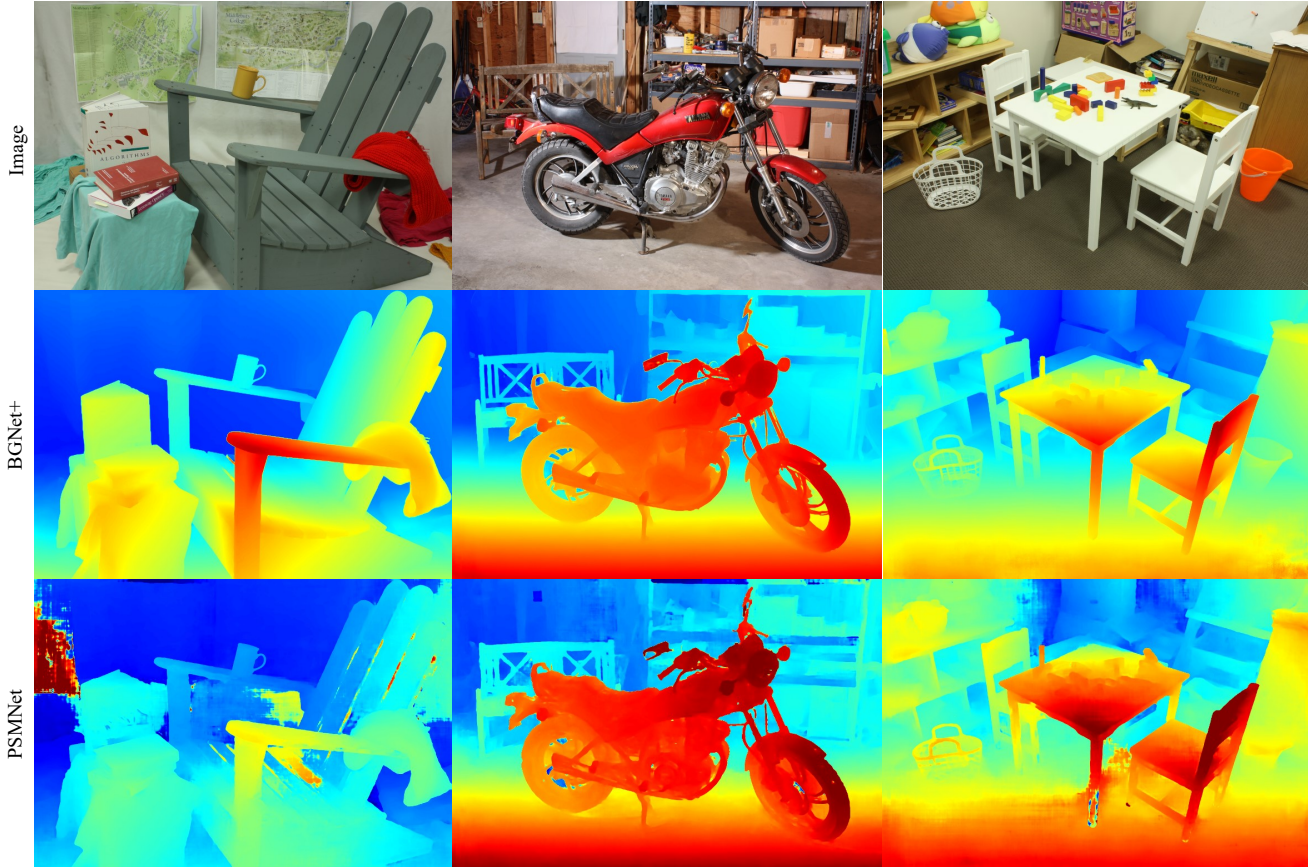


Figure 5. Qualitative results of generalization performance evaluation on the Middlebury 2014 dataset. The first row shows the left RGB images of the stereo pairs, the second row shows the results of BGNet+, and the third row shows the results of PSMNet. The models were trained only on the synthetic SceneFlow dataset.

Network	Bad 2.0 (%)
PSMNet [5]	25.1
iRestNet-i2 [18]	19.8
GANet [44]	20.3
DSMNet [45]	13.8
DeepPruner-Fast [9]	17.8
DeepPruner-Fast-BG	16.5
BGNet	17.5
BGNet+	17.2
BGNet (IRS)	13.5
BGNet+ (IRS)	11.6

Table 6. Generalization ability on the Middlebury 2014 dataset of half resolution. The percentage of pixels with errors larger than 2 pixels (Bad 2.0) is reported. All the models are trained on the synthetic datasets.

5. Conclusion

In this paper, we have proposed a cost volume upsampling module in the learned bilateral grid. The upsam-

pling module can be seamlessly embedded into many existing end-to-end network architectures and can accelerate these networks significantly while maintaining comparable accuracy. In addition, based on this upsampling module, we design two real-time networks, BGNet and BGNet+. They outperform all the published networks with computational time less than 50 ms on the KITTI 2012 and KITTI 2015 datasets. The experimental results also demonstrate the good generalization ability of our network.

In the future, we plan to apply our approach to both monocular depth estimation and depth completion tasks.

Acknowledgements. This work was supported by Orbbe Inc. (No. W2020JSKF0547), the National Natural Science Foundation of China (No. U20A20185, 61972435, 62076086), the Natural Science Foundation of Guangdong Province (2019A1515011271), and the Shenzhen Science and Technology Program (No. RCYX20200714114641140, JCYJ20190807152209394).

References

- [1] Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: a large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, 63(11):1–11, 2020. [1](#)
- [2] Jonathan T Barron, Andrew Adams, YiChang Shih, and Carlos Hernández. Fast bilateral-space stereo for synthetic defocus. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4466–4474, 2015. [2](#)
- [3] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Machine Vision Conference*, volume 11, pages 1–11, 2011. [2](#)
- [4] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698, 1986. [5](#)
- [5] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#), [8](#)
- [6] Jiawen Chen, Andrew Adams, Neal Wadhwa, and Samuel W Hasinoff. Bilateral guided upsampling. *ACM Transactions on Graphics*, 35(6):1–8, 2016. [2](#)
- [7] Jiawen Chen, Sylvain Paris, and Frédo Durand. Real-time edge-aware image processing with the bilateral grid. *ACM Transactions on Graphics*, 26(3):103–es, 2007. [1](#), [2](#), [3](#)
- [8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015. [2](#)
- [9] 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*, pages 4384–4393, 2019. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. [1](#), [4](#), [7](#)
- [11] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, 36(4):1–12, 2017. [1](#), [2](#)
- [12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. [2](#)
- [13] Xiaoyang Guo, Kai Yang, Wukui Yang, Xiaogang Wang, and Hongsheng Li. Group-wise correlation stereo network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3273–3282, 2019. [3](#), [4](#)
- [14] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. [1](#), [3](#), [4](#), [6](#), [7](#)
- [15] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. [2](#)
- [16] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for edge-aware depth prediction. In *European Conference on Computer Vision*, 2018. [1](#), [2](#), [7](#)
- [17] Hyunmin Lee and Yongho Shin. Real-time stereo matching network with high accuracy. In *IEEE International Conference on Image Processing*, pages 4280–4284. IEEE, 2019. [7](#)
- [18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. [1](#), [2](#), [6](#), [7](#), [8](#)
- [19] Zhengfa Liang, Yulan Guo, Yiliu Feng, Wei Chen, Linbo Qiao, Li Zhou, Jianfeng Zhang, and Hengzhu Liu. Stereo matching using multi-level cost volume and multi-scale feature constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. [2](#)
- [20] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. [2](#), [6](#), [7](#)
- [21] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. [7](#)
- [22] Xing Mei, Xun Sun, Mingcai Zhou, Shaohui Jiao, Haitao Wang, and Xiaopeng Zhang. On building an accurate stereo matching system on graphics hardware. In *IEEE International Conference on Computer Vision Workshops*, pages 467–474. IEEE, 2011. [2](#)
- [23] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2, 2015. [1](#), [4](#), [7](#)
- [24] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. [2](#)
- [25] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. [5](#)

- [26] Christian Richardt, Douglas Orr, Ian Davies, Antonio Criminisi, and Neil A Dodgson. Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid. In *Proceedings of the European Conference on Computer Vision*, pages 510–523. Springer, 2010. [2](#)
- [27] 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*, pages 31–42. Springer, 2014. [1](#)
- [28] 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*, pages 31–42. Springer, 2014. [5](#)
- [29] Akihito Seki and Marc Pollefeys. Sgm-nets: Semi-global matching with neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 231–240, 2017. [7](#)
- [30] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019. [5](#)
- [31] Xiao Song, Xu Zhao, Liangji Fang, Hanwen Hu, and Yizhou Yu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. *International Journal of Computer Vision*, pages 1–21, 2020. [2](#), [7](#)
- [32] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *IEEE International Conference on Computer Vision*, pages 839–846. IEEE, 1998. [1](#)
- [33] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. [2](#)
- [34] Longguang Wang, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, Wei An, and Yulan Guo. Learning parallax attention for stereo image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12250–12259, 2019. [2](#)
- [35] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. Fadnet: A fast and accurate network for disparity estimation. *International Conference on Robotics and Automation*, 2020. [2](#), [6](#), [7](#)
- [36] Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large synthetic indoor robotics stereo dataset for disparity and surface normal estimation. *arXiv preprint arXiv:1912.09678*, 2019. [7](#)
- [37] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1959–1968, 2020. [1](#), [2](#), [6](#), [7](#)
- [38] 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*, pages 5515–5524, 2019. [5](#)
- [39] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic information for disparity estimation. In *Proceedings of the European Conference on Computer Vision*, pages 636–651, 2018. [2](#)
- [40] Qingxiong Yang, Narendra Ahuja, Ruigang Yang, Kar-Han Tan, James Davis, Bruce Culbertson, John Apostolopoulos, and Gang Wang. Fusion of median and bilateral filtering for range image upsampling. *IEEE Transactions on Image Processing*, 22(12):4841–4852, 2013. [1](#)
- [41] Kuk-Jin Yoon and In So Kweon. Adaptive support-weight approach for correspondence search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):650–656, 2006. [1](#)
- [42] Jure Zbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1592–1599, 2015. [2](#)
- [43] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *The Journal of Machine Learning Research*, 17(1):2287–2318, 2016. [1](#), [7](#)
- [44] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. [1](#), [3](#), [6](#), [7](#), [8](#)
- [45] Feihu Zhang, Xiaojuan Qi, Ruigang Yang, Victor Prisacariu, Benjamin Wah, and Philip Torr. Domain-invariant stereo matching networks. *arXiv preprint arXiv:1911.13287*, 2019. [7](#), [8](#)