EDNet: Efficient Disparity Estimation with Cost Volume Combination and Attention-based Spatial Residual

Songyan Zhang[1], Zhicheng Wang[1]*, Qiang Wang[2], Jinshuo Zhang[1], Gang Wei[1], Xiaowen Chu[2]

[1]CAD Research Center, Tongji University

[2]Department of Computer Science, Hong Kong Baptist University

{spyder, zhichengwang, zhangjinshuo, weigang}@tongji.edu.cn

{qiangwang, chxw}@comp.hkbu.edu.hk

Abstract

Existing state-of-the-art disparity estimation works mostly leverage the 4D concatenation volume and construct a very deep 3D convolution neural network (CNN) for disparity regression, which is inefficient due to the high memory consumption and slow inference speed. In this paper, we propose a network named EDNet for efficient disparity estimation. Firstly, we construct a combined volume which incorporates contextual information from the squeezed concatenation volume and feature similarity measurement from the correlation volume. The combined volume can be next aggregated by 2D convolutions which are faster and require less memory than 3D convolutions. Secondly, we propose an attention-based spatial residual module to generate attention-aware residual features. The attention mechanism is applied to provide intuitive spatial evidence about inaccurate regions with the help of error maps at multiple scales and thus improve the residual learning efficiency. Extensive experiments on the Scene Flow and KITTI datasets show that EDNet outperforms the previous 3D CNN based works and achieves state-of-the-art performance with significantly faster speed and less memory consumption.

1. Introduction

Accurate and fast depth estimation is of great significance to many applications like robot navigation, 3D reconstruction and autonomous driving. Instead of depth regression from a single-view RGB image, stereo matching is to conduct correspondence analysis between pixels of stereo images and compute the disparity d for each pixel. Depth can be then calculated by $(\frac{fB}{d})$, where f is the camera's focal length and B is the distance between two camera centers, also called baseline in stereo vision.

While traditional methods based on hand-crafted feature extraction and matching cost aggregation tend to fail on those textureless and repetitive regions in the images, convolutional neural networks (CNNs) have been widely adopted to conquer those difficulties in stereo matching. Several recent methods [19, 3, 13, 42] have achieved state-of-the-art performance by constructing a 4D concatenation volume which follows 3D convolution blocks for aggregation. Although the 4D concatenation cost volume can preserve the rich contextual information in conjunction with the strong regularization ability of 3D convolutions, it significantly increases the computation cost and usually cannot perform real-time disparity inference. Moreover, the concatenation volume incorporates no feature similarity measurement, which means that the model has to learn correspondence from scratch. Besides, DispNetC [22] formed a low-cost correlation layer with 2D convolutions to conduct correspondence analysis between the left and right feature maps. The following works [21, 38, 26] adopted the similar method as it keeps a good balance between speed and accuracy. However, as the correlation map is produced with only one single feature channel for each disparity level, the performance is less competitive. Thus, this raises the question of how to make full use of the complementary advantages of the concatenation volume and correlation volume.

Since ResNet [15] has revealed that the residual convolution block can improve the training efficiency by learning a residual mapping instead of the desired underlying one, it has been adopted as a popular approach to refine the disparity estimation [34, 31, 21, 37]. To be specific in stereo matching, learning an additive correction to the coarse disparity map is easier and more efficient than directly learning the fine-grained one. However, some works failed to provide the residual learning module with the fitting error information [33, 40], or computed the estimated error at only one scale [34, 26, 17] but adopted it to learn the disparity maps at multiple scales. The error map from one single
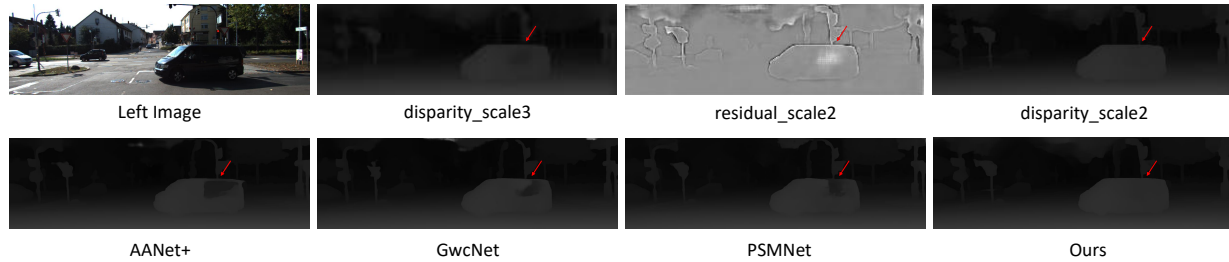
Figure 1: The first row is the visualization of residual learning process from scale 3 to scale 2. The residual_scale2 is learned from the disparity_scale3 for correcting the disparity_scale2. With our proposed modules, sharp edges and overall structures can be recovered. Other state-of-the-art methods fail to generate the accurate disparity estimation in low-texture regions as shown in the second row. Please pay more attention to regions pointed by the red arrow.

scale cannot provide the precise error information at other scales, which makes the residual learning method less effective. Furthermore, even if the error map is provided at each corresponding scale [31], the conventional residual learning method has no explicit spatial guidance about where to intervene. As the regions with inaccurate estimation deserve more attention, we argue that residual learning could be more efficient if the spatial attention about the learning errors is provided.

To address the above issues, we propose EDNet which is composed of a combined volume to generate robust feature representations, and an attention-based residual module to learn the disparity refinement. Firstly, the proposed combined volume alleviates the information loss by employing the squeezed concatenation volume and preserves the feature similarity measurement with the correlation volume. We then adopt 2D convolutions for further aggregation so that the significant memory consumption and computational complexity of 3D CNNs can be avoided. Secondly, inspired by the attention mechanism, we adopt a spatial attention module to generate the attention-aware residual features. Therefore, the residual learning module can have intuitive spatial evidence about inaccurate regions to compute a specific correction. We follow the coarse-to-fine strategy and compute the attention-aware residuals across different scales. With the error maps provided at each scale, the residual module can learn a corresponding correction accordingly and improve the learning efficiency. As shown in Figure 1, our network can generate an accurate and continuous disparity map even in low-texture regions. The contributions of our work can be summarized as follows:

- We propose a low-cost but effective method to aggregate the 3D correlation features and 4D concatenation volume together by constructing a combined volume, which can be further processed by fast 2D convolutions. Compared with others, our correspondence analysis can preserve both the contextual information and feature similarities even with 2D convolutions.

- We design an Attention-based Residual (AR) module to learn the disparity refinement at each scale. In the AR module, the attention mechanism is applied to the concatenated maps of RGB image, estimated disparity and estimated error to improve the learning efficiency.

- Compared to those existing methods based on 3D CNNs, our proposed EDNet achieves state-of-the-art accuracy on the public Scene Flow [22] and KITTI [24, 10] datasets with less than 25% running memory requirement compared with Bi3D [1] and is 45× faster than GANet [42].

The rest of the paper is organized as follows. We introduce some related studies about stereo matching based on CNNs in Section 2. Section 3 introduces the methodology and implementation of our proposed EDNet. We demonstrate our experimental results in Section 4. We finally conclude the paper in Section 5.

2. Related Works

A classical stereo matching pipeline consists of four steps [30]. In recent years, CNNs have drawn great attention and been introduced to tackle the stereo matching task [41]. In this section, we briefly discuss those common mechanisms for computing the matching cost with CNNs and review the approaches with the residual learning method.

2.1. Matching Cost Computation

CNN based matching cost computation methods make a great contribution to the stereo matching accuracy. There are two popular approaches for matching cost computation. The first one is using either a layer of 2D [7] or 1D [22] convolutional operations, called correlation layer. Such an inner product between feature vectors is adopted in [4, 9, 21, 38]. Liang et al. [21] build a correlation volume for initial disparity estimation which follows a disparity refinement module by learning through feature consistency. Wang et al. [34, 35] make some modification and propose

a point-wise correlation volume to preserve fast computation. Another popular method to compute matching cost is to form a 4D volume by concatenating the corresponding features from the opposite stereo images across each disparity level. 3D convolutions are followed to aggregate features and regress disparity. This method can be found first in [19]. Chang *et al.* [3] improve Kendall's approach [19] by designing a spatial pyramid pooling module [14] so that correspondence estimation can benefit from the image features with rich object context information. Guo *et al.* [13] combine the concatenation volume with the group-wise correlation volume and improve the accuracy with 3D convolutions. The best performance on Scene Flow dataset comes from [25] which introduces the idea of DenseNet [16] to further improve PSMNet [3]. Zhang *et al.* propose GANet [42] with two guided aggregation layers and fifteen 3D convolutions to achieve state-of-the-art performance. Cheng *et al.* improves the stereo matching based on 3D convolutions with the help of NAS [5].

2.2. Residual Learning for Stereo Matching

The residual learning concept is proposed by He *et al.* [15] which turns to be an efficient way to train a CNN model and has been adopted by many works. In the stereo matching task, the residual learning strategy is widely used for refining disparity estimation [21, 34, 20, 37, 11, 2, 18]. Pang *et al.* [26] present a cascade residual learning scheme and adopt a two-stage CNN, in which the second stage refines the estimation by producing residual signals. Stucker *et al.* [32] specially build a U-Net [28] based network to enhance the reconstruction by regressing a residual correction. In order to meet the need for real-time inference, [36] takes residual learning strategy to flexibly output disparity estimation according to the requirement of applications. Song *et al.* [31] manage to aggregate edge information for residual learning and thus construct a multi-task network for edge detection and stereo matching.

3. Methodology

3.1. Network Architecture

The architecture of our proposed EDNet is shown in Figure 2. We exploit the structure of DispNetC [22] as the backbone with extensive modifications. For feature extraction, the last left and right feature maps of conv3 from the weights-sharing encoder network are used to form the combined volume which is composed of a squeezed concatenation volume and a correlation volume which will be discussed in Section 3.2. The detail of our proposed cost volume combination method can be found in the left bottom corner of Figure 2. 2D convolutions are then employed to aggregate the combined volume and regress the disparity. In the decoder part, we follow the coarse-to-fine strategy

to refine the disparity progressively. The spatial attention module is applied in order to generate attention-aware residual features, which will be introduced in Section 3.3. The stacked hourglass module in PSMNet [3] is used for residual regression but is implemented by 2D convolutions. The attention-based spatial residual module is well illustrated in the right bottom corner of Figure 2. Different from DispNetC [22] which has 6 scales of output, we reduce the disparity prediction to 4 scales, removing the prediction at 1/16 and 1/32 of full resolution.

3.2. Cost Volume Combination

Previous works simply build a correlation volume [22, 26, 34] or a 4D concatenation volume [19, 13, 3] which follows 2D or 3D convolutions for aggregation. However, a single cost volume cannot meet the need of preserving contextual information and feature similarity at the same time. GwcNet [13] improves the performance by combining the group-wise correlation and concatenation volume, but 3D convolutions are required for aggregation which leads to higher memory consumption and more complex computation. To this end, we propose to combine the correlation volume and 4D concatenation volume in a more efficient way aiming at taking advantages of both two cost volumes.

Given a pair of stereo features $\mathbf{f}_L$ and $\mathbf{f}_R$, we follow the 1D-correlation in DispNetC [22] to calculate the correspondence at each disparity level d. The correlation volume is computed as:

$$\mathbf{C}_{corr}(d, x, y) = \frac{1}{N} < \mathbf{f}_L(x - d, y), \mathbf{f}_R(x, y) > \quad (1)$$

where $< x_1, x_2 >$ is the inner product of two feature matrices x_1 and x_2, and N is the channel number of input features. The shape of the correlation volume is $N \times D \times H \times W$, where N denotes the batch size, D is the estimated disparity range and the spatial size is $H \times W$. Then we construct the 4D concatenation volume by concatenating the left and right feature maps, i.e.,

$$\mathbf{C}_{concat}(d, x, y) = Concat\{\mathbf{f}_L(x - d, y), \mathbf{f}_R(x, y)\} \quad (2)$$

After obtaining the concatenation volume with the shape $N \times D \times C \times H \times W$, we use three 3D convolutions for aggregation and compress it into 1 channel. The aggregated concatenation volume now has the shape $N \times D \times 1 \times H \times W$. It is then squeezed into $N \times D \times H \times W$, the same shape as the correlation volume. The correlation volume and squeezed concatenation volume are finally concatenated to form the combined volume. In this way, both the contextual information and feature similarity measurement are incorporated in the combined volume. Further aggregation can be done by 2D convolutions instead of 3D convolutions, which are more efficient. The final combined
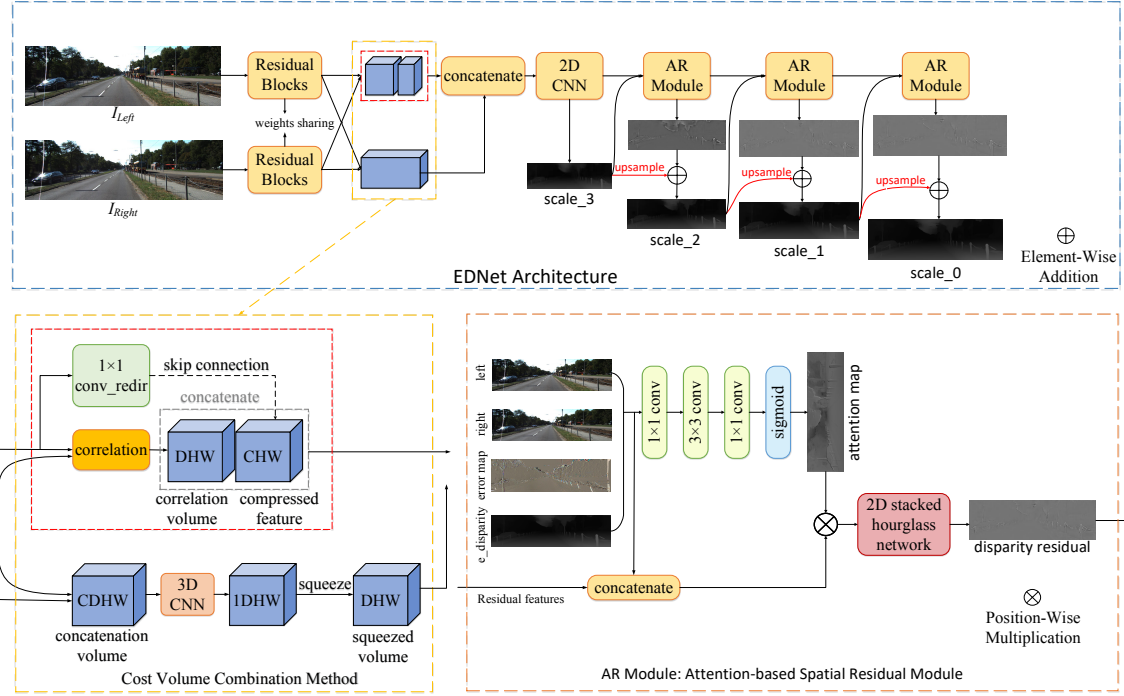
Figure 2: An overview of our proposed EDNet. We exploit the architecture of DispNetC as the backbone. The attention-based spatial residual module and combined volume are proposed for accuracy and efficiency improvement. To provide better visualization, the skip connections in DispNetC and some other data flow are omitted here.

volume is formed as:

$$\mathbf{C}_{comb}(x,y) = Concat\{\mathbf{C}_{corr}(x,y), \mathbf{C}_{concat}(x,y)\} \quad (3)$$

3.3. Attention-based Spatial Residual

The normal residual learning method lacks the spatial evidence about where the errors occur. We propose an attention-based spatial residual module to guide the residual learning process to pay more attention to those inaccurate regions across the whole spatial space. According to the estimated disparity $\hat{d}^s$ at scale s, a synthesized left image $\tilde{I}_L^s$ can be obtained by warping the right image I_R^s, i.e.,

$$\tilde{I}_L^s(x,y) = I_R^s(x + \hat{d}^s(x,y), y) \quad (4)$$

With the warped left image and target left image, we can get the error $E_L^s = |\tilde{I}_L^s - I_L^s|$. A spatial attention module with 3 layers of 2D convolution which are 1×1, 3×3 and 1×1 respectively is applied. The spatial attention feature map $\mathbf{f}_a^s$ is compressed into one channel followed by the sigmoid function to compute the spatial attention matrix whose size is $N \times 1 \times H \times W$. The input of error map and color stereo images enables the spatial attention module to learn an attention distribution on blurry object boundaries and mismatched pixels. Akin to FADNet [34], CRL[34], FlowNet2 [17], the input to spatial attention module is the

concatenation of the stereo images, error map and estimated disparity map.

We follow DispNetC [22] to preserve both the high-level information and local information by skip-connection. The concatenation of 'upconvolution' feature maps from the decoder network and corresponding feature maps from the encoder network is then concatenated with the input of attention module to form the residual feature maps $\mathbf{f}_r^s \in R^{H \times W \times C}$. The final attention-aware residual features $\mathbf{f}_{ar}^s \in R^{H \times W \times C}$ at scale s are computed by multiplying the attention matrix and residual features $\mathbf{f}_r^s$, i.e.,

$$\mathbf{f}_{ar}^s = \mathbf{f}_{ar}^s \otimes \sigma(\mathbf{f}_a^s) \quad (5)$$

where $\sigma(\cdot)$ denotes the sigmoid function. The attention-aware residual features are then input to the stacked hourglass module for residual regression. The stacked hourglass module has the same encoder-decoder structure as PSMNet [3] but is implemented with 2D convolutions.

The attention-based spatial residual module is repeated 3 times as we increase the resolution of the disparity map progressively. Different from the aforementioned works [34, 26], we compute error maps across multiple scales as the error information changes accordingly after the correction. Therefore, we provide the residual learning module with constantly updated error maps across multiple scales instead of a single error map at full resolution.

3.4. Multi-scale Residual Learning

Instead of building a cascade architecture with residual refinement at the second stage [17, 26, 34], we simply replace the direct disparity estimation with a residual estimation over all scales except the smallest S at which the initial disparity is computed. The multi-scale residual outputs are denoted as $\{r^s\}_{s=0}^{S-1}$ where 0 represents the scale of full resolution. For the rest S scales, the estimated disparity at the previous scale is first upsampled to the current scale $\hat{d}_{up}^s$ using bilinear interpolation function and then added to the residual for refinement. The final predicted disparity $\hat{d}^s$ at scale s is produced as:

$$\hat{d}^s = \hat{d}_{up}^s + r^s, 0 \leq s \leq S - 1 \qquad (6)$$

3.5. Loss Function

Given the output disparity map at different scales, we adopt the pixel-wise smooth L1 loss to train our EDNet at scale s:

$$L^s(d^s, \hat{d}^s) = \frac{1}{N} \sum_{i=1}^{N} smooth_{L_1}(d_i^s - \hat{d}_i^s), \qquad (7)$$

where N is the number of pixels of the disparity map, $\hat{d}_i^s$ is the ith element of the predicted disparity $\hat{d}^s$, and d^s represents the ground truth disparity. The smooth L1 loss function is:

$$smooth_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise.} \end{cases} \qquad (8)$$

The final loss function is a combination of losses over all scales, i.e.,

$$L = \sum_{s=0}^{S} \lambda^s L^s(d^s, \hat{d}^s) \qquad (9)$$

where λ^s is a scalar for adjusting the loss weight at scale s.

4. Performance Evaluation

4.1. Datasets and Evaluation Metrics

Three public datasets are used for training and testing our EDNet. The Scene Flow dataset [22] consists of 39,824 pairs of synthetic stereo RGB images (35,454 pairs for training and 4,370 pairs for testing) with a full resolution of 960×540. Both KITTI 2012 [10] and KITTI 2015 [24] are datasets of real scenes with a full resolution of 1242×375. The ground truth of these two datasets is generated by lidar so that only sparse ground truth is available. We evaluate our model on Scene Flow dataset with the end-point error (EPE), 1-pixel error and 3-pixel error. The end-point error computes the mean disparity error in pixels while the 1-pixel error and 3-pixel error measure the average percentage

of the pixel whose EPE is bigger than 1 pixel and 3 pixels respectively. The official metrics (e.g., D1-all) are reported for evaluation on KITTI 2012 and KITTI 2015 datasets.

4.2. Implementation Details

We implemented our EDNet by PyTorch [27] and trained the model with Adam (momentum=0.9, beta=0.999) as optimizer. For Scene Flow dataset, raw images are randomly cropped to 320×640 as input. Our training is performed on 2 NVIDIA RTX 2080Ti GPUs for 70 epochs with a batch size of 8 (4 on each GPU). We follow the training strategy in AANet [37], where the initial learning rate is set to 0.001 and decreased by half every 10 epochs after 20th epoch. The loss weights are set to $\lambda_0 = 1.0, \lambda_1 = \lambda_2 = 0.8, \lambda_3 = 0.6$. The crop size for KITTI is set as 256×512. Due to the insufficient training samples of both KITTI 2012 and KITTI 2015, the pre-trained Scene Flow model is used for fine-tuning on the mixed KITTI 2012 and KITTI 2015 training sets for the first 1000 epochs which follows another 400 epochs of training on two datasets to get the submission result respectively. We use a constant learning rate of 0.0001 for KITTI datasets. Inspired by [23] that searching the correspondence at a coarse scale can be beneficial, especially in low-texture or textureless regions, the loss weights are set as $\lambda_0 = 0.6, \lambda_1 = \lambda_2 = 0.8, \lambda_3 = 1.0$. For all datasets, color normalization is taken into use with the mean ([0.485, 0.456, 0.406]) and variation ([0.229, 0.224, 0.225]) of the ImageNet [6] for data pre-processing. The maximum disparity is set as 192 pixels.

4.3. Ablation Study

To validate the effectiveness of each component in EDNet, we evaluate our model with different configurations on Scene Flow dataset. All the experimental results are obtained after 10 epochs of training. As shown in Table 1, removing either of the combined volume or attention-based spatial residual leads to a clear drop of performance.

Cost Volume Combination: As shown in Figure 3, models without the combined volume suffer from inaccurate and discontinuous disparity estimation, especially in low-texture regions. The possible reason might be that a single cost volume is unable to avoid the information loss which leads to less robust feature representations. As shown in Table 1, the EPE decreases from EDNet-NS's 1.07 to EDNet-F's 1.00. The comparison among EDNet-NRS, EDNet-NRCo and EDNet-NR can validate the effectiveness of our proposed combined volume as well.

Attention-based Spatial Residual: It can be directly analyzed from EDNet-NR and EDNet-NA in Table 1 that the residual learning module brings significant improvement in terms of accuracy, about 36% decrease of EPE. The comparison between EDNet-NA and EDNet-F in Table 1 demonstrates that the learning efficiency can be further im-

Method	Cost Volume		Residual Module		Results		
	correlation	s-concatenation	normal	attention	EPE	>1px(%)	>3px(%)
EDNet-NRS	✓				1.67	27.9	9.7
EDNet-NRCo		✓			1.89	29.9	10.7
EDNet-NR	✓	✓			1.63	27.4	9.8
EDNet-NA	✓	✓	✓		1.04	12.7	**5.4**
EDNet-NS	✓			✓	1.07	13.1	5.8
EDNet-F	✓	✓		✓	**1.00**	**12.2**	**5.4**

Table 1: Evaluation of EDNet with different settings. We compute the end-point error, 1-pixel and 3-pixel error on Scene Flow dataset. We use "Co", "S", "R", "A" to denote the correlation volume, squeezed concatenation volume, normal residual, and attention-based residual respectively. "N" stands for "not applied". "EDNet-F" represents the model with all our proposed components and "s-concatenation" means the squeezed concatenation volume.
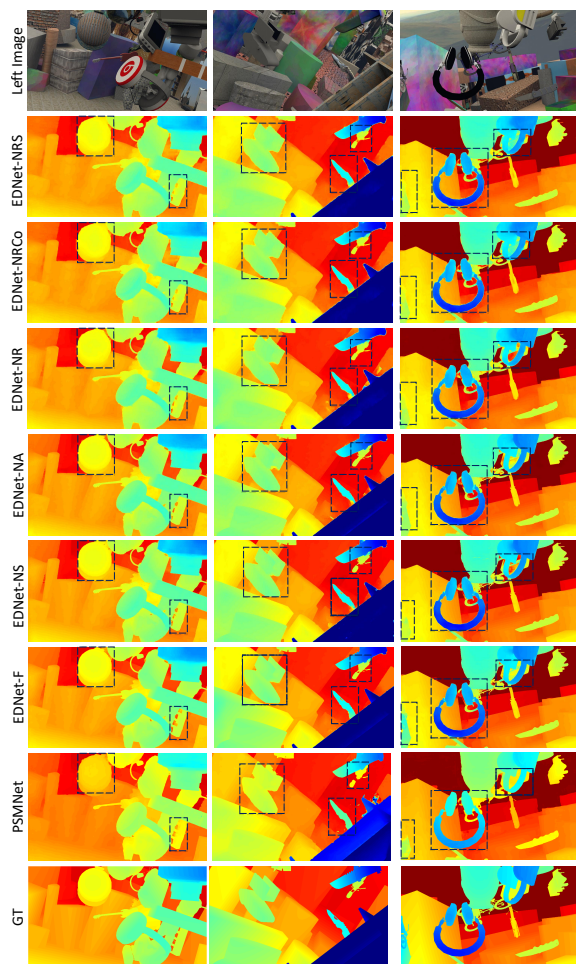

Figure 3: Visualization of the ablation study results on Scene Flow dataset. Our proposed combined volume and spatial residual module can bring great performance improvement especially at bounding box areas.

proved with the help of the attention mechanism. The visualization of EDNet-NA and EDNet-F shown in Figure 3 illustrates that more details like the shaper object boundaries
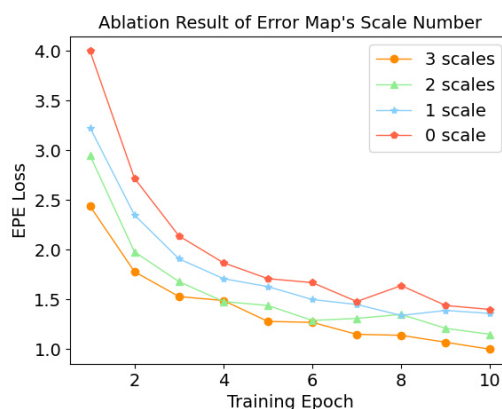

Figure 4: Comparison among models with different numbers of error maps. The multi-scale error maps speed up the convergence and achieve a lower loss. "n scale(s)" indicates that the error map is applied to the residual features from scale_0 to scale_n.

can be recovered with the attention-based spatial residual.

Multi-scale Error Maps: Experiments are conducted to prove that our multi-scale error maps mechanism is of great importance to the residual learning. We remove the error map from the residual features as well as the input of attention module from scale 2 to all scale. As shown in Figure 4, models with error maps at multiple scales can achieve a lower EPE loss with a faster convergence speed. Such performance gain comes at a low cost of extra computation. Multi-scale error maps enable the residual module to learn from the error at the corresponding scale and thus greatly explore the ability of residual learning.

4.4. Experimental Results

In this subsection, we compare our method with those existing state-of-the-art methods in the aspect of inference speed, memory consumption and accuracy on Scene Flow, KITTI 2015 and KITTI 2012 datasets. We also evaluate the model generalization on Middlebury 2014 dataset [29].

Method	PSMNet [3]	GANet [42]	GwcNet [13]	Bi3D [1]	DispNetC [22]	AANet+ [37]	Ours
EPE	1.09	0.84	0.76	0.73	1.68	<u>0.72</u>	**0.63**
Time (s)	0.453	3.302	0.254	OOM	**0.025**	0.068	<u>0.059</u>

Table 2: EPE values on Scene Flow dataset of several state-of-the-art methods. Our method achieves the best scores and has a competitive inference speed. The inference time is tested on a single NVIDIA 2080Ti GPU at the resolution of 576×960 for a fair comparison. OOM denotes out of memory. "<u>X</u>" indicates the second best.
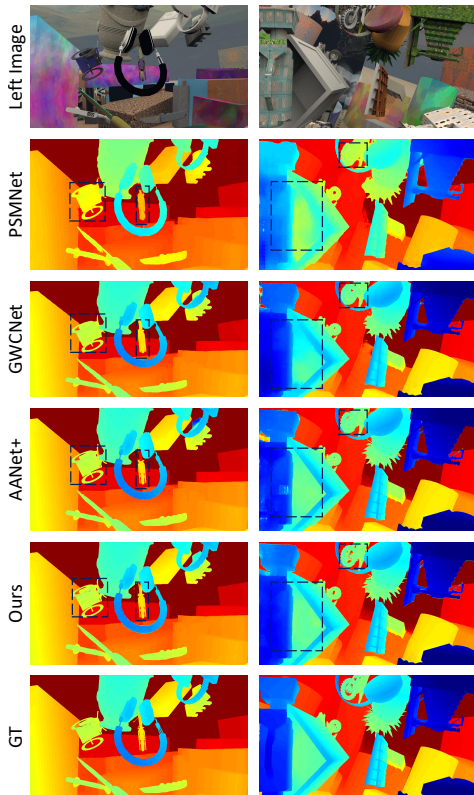

Figure 5: Visual comparison of our EDNet and some other top-performing methods. Please enlarge the bounding box areas for more details.

Method	Noc (%)		All (%)		Time (s)
	fg	all	fg	all	
GANet [42]	3.37	**1.73**	3.82	**1.93**	0.36
GCNet [19]	5.58	2.61	6.16	2.87	0.9
PSMNet [3]	4.31	2.14	4.62	2.32	0.41
GwcNet [13]	3.49	1.92	3.93	2.11	0.32
SegStereo [38]	3.70	2.08	4.07	2.25	0.6
MC-CNN [43]	7.64	3.33	8.88	3.89	67
HD^3 [39]	3.43	1.87	3.63	2.02	0.14
CSN [12]	3.55	1.78	4.03	2.00	0.6
DeepPruner-B [8]	3.18	1.95	3.56	2.15	0.18
Bi3D [1]	**3.11**	2.01	**3.48**	2.21	0.48
Ours	3.33	2.31	3.88	2.53	**0.05**
AANet [37]	4.93	2.32	5.39	2.55	0.06
DeepPruner-F [8]	3.43	2.35	3.91	2.59	0.06
DispNetC [22]	3.72	4.05	4.41	4.34	0.06
FADNet [34]	**3.07**	2.59	**3.50**	2.82	0.05
MADNet [33]	8.41	4.27	9.20	4.66	**0.02**
Ours	3.33	**2.31**	3.88	**2.53**	0.05

Table 3: Benchmark results on KITTI 2015 test sets. "Noc" and "All" indicate the percentage of outliers averaged over ground truth pixels of non-occluded and all regions respectively. "fg" and "all" indicate the percentage of outliers averaged over the foreground and all ground truth pixels respectively. "**X**" indicates the best result.

Scene Flow Dataset: As shown in Table 2, our proposed EDNet not only outperforms all the competing state-of-the-art methods with the lowest EPE but also runs significantly faster: approximately 7.5× faster than PSMNet [3], 4× faster than GwcNet [13], 55× faster than GANet [42]. Compared with DispNetC [22] and StereoNet [20], EDNet improves the performance by 60% and 40% respectively. Figure 5 gives a visual comparison between our EDNet and other outstanding methods. Sharper object boundaries and more continuous disparity maps can be generated by ED-Net, indicating the value of our proposed approach.

KITTI Datasets: We divide the comparison into two groups as shown in Table 3 and Table 4. First, compared with other top-performing methods, our EDNet still achieves competitive results when evaluating non-occluded

pixels while runs considerably faster according to the benchmark. Then we compare our EDNet with previous real-time models. Experimental results from Table 3 and Table 4 show that our work can produce competitive estimation. To stress the efficiency of our proposed method, we compare the computational complexity, memory consumption as well as the inference speed with some popular 3D convolutions based models. Table 5 comprehensively shows that our model requires less memory consumption: approximately 50% less than PSMNet [3], 40% less than GwcNet [13], 60% less than GANet [42] and 70% less than Bi3D [1]. Moreover, our EDNet has lower computational complexity as our work has the lowest FLOPs (the lower the better) and runs significantly faster: approximately 7× faster than PSMNet [3], 17× faster than Bi3D [1] and 45× faster than GANet [42]. Figure 6 visualizes the disparity and error maps of our method and other state-of-the-art works on KITTI 2015 dataset.
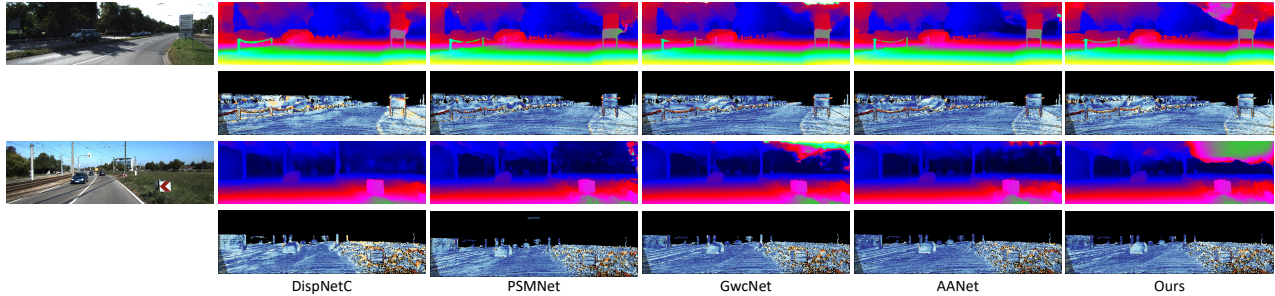
Figure 6: Results of the disparity prediction for KITTI 2015 testing data. The leftmost column shows left images of the stereo pairs. The rest five columns show the disparity maps predicted by DispNetC [22], PSMNet [3], GwcNet [13], AANet[37] and our EDNet, as well as their error maps.

Method	Out (%)		Avg		Time (s)
	Noc	All	Noc	All	
GANet [42]	2.18	2.79	**0.5**	**0.5**	0.36
GCNet [19]	2.71	3.46	0.6	0.7	0.90
PSMNet [3]	2.44	3.01	**0.5**	0.6	0.41
GwcNet [13]	**2.16**	**2.71**	**0.5**	**0.5**	0.32
SegStereo [38]	2.66	3.19	**0.5**	0.6	0.60
MC-CNN [43]	3.90	5.45	0.7	0.9	67
Ours	2.97	3.67	**0.5**	0.6	**0.05**
AANet [37]	**2.90**	**3.60**	**0.5**	**0.6**	0.06
StereoNet [20]	4.91	6.02	0.8	0.9	**0.015**
DispNetC [22]	7.38	8.11	0.9	1.0	0.06
FADNet [34]	3.98	4.61	0.6	0.7	0.05
Ours	2.97	3.67	**0.5**	**0.6**	0.05

Table 4: Benchmark results on KITTI2012 test sets. Both the percentage of pixels with error bigger than 2 and the overall EPE are reported over non-occluded and all regions.

Method	Memory (GB)	GFLOPs	Time (s)
PSMNet [3]	4.83	937.9	0.393
GANet [42]	6.53	1936.98	2.43
GwcNet [13]	4.27	899.99	0.272
Bi3D [1]	10.74	4212.05	0.899
Ours	**2.52**	**162.92**	**0.053**

Table 5: Comparisons of the runtime, running memory and computational cost. All the results are tested on a single NVIDIA RTX 2080Ti GPU at a resolution of 1248×384.

Middlebury 2014 Dataset: We evaluate the model generalization on Middlebury 2014 dataset without additional training. All the models are pretrained on Scene Flow and finetuned on KITTI. As observed from Figure 7, the performance of our EDNet is superior to others, as more smooth and continuous disparity estimation in low-texture regions can be produced. Moreover, our EDNet can preserve better overall object structures and generate sharper edges.
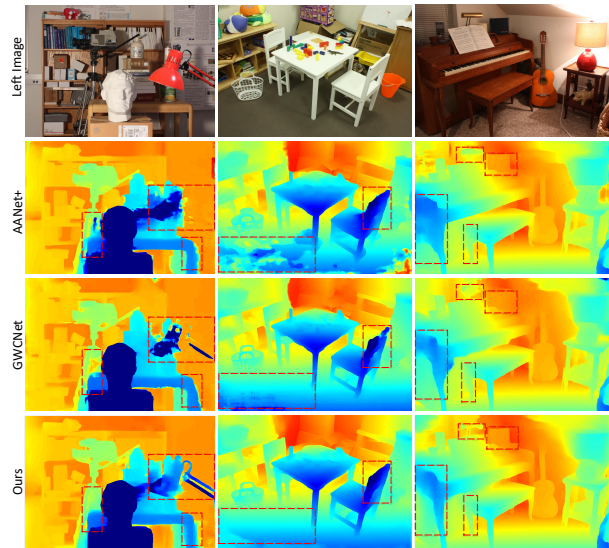

Figure 7: Evaluation of model generalization on Middlebury 2014 dataset. Our EDNet produces more smooth and continous disparity maps. Sharper object boundaries as well as better overall structures can be recovered. Please zoom in bounding box areas for a further comparison.

5. Conclusion

We have exploited an efficient architecture called EDNet with the proposed combined volume and attention-based residual module. We show that the combination of the correlation volume and squeezed 4D concatenation volume is of great importance to robust feature representations, especially in ill-posed regions. Besides, the employment of the spatial attention module greatly improves the efficiency of residual learning with multi-scale error maps. Extensive experimental results on KITTI and Scene Flow datasets have demonstrated the superiority of our method when comparing with previous state-of-the-art methods. The future work would be applying our work to other depth-related tasks, e.g., 3D reconstruction and robot navigation.

References

[1] A. Badki, A. Troccoli, K. Kim, J. Kautz, P. Sen, and O. Gallo. Bi3d: Stereo depth estimation via binary classifications. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1597–1605, 2020. 2, 7, 8

[2] K. Batsos and P. Mordohai. Recresnet: A recurrent residual cnn architecture for disparity map enhancement. In *2018 International Conference on 3D Vision (3DV)*, pages 238–247, 2018. 3

[3] J. Chang and Y. Chen. Pyramid stereo matching network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1, 3, 4, 7, 8

[4] Z. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang. A deep visual correspondence embedding model for stereo matching costs. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 972–980, 2015. 2

[5] Xuelian Cheng, Yiran Zhong, Mehrtash Harandi, Yuchao Dai, Xiaojun Chang, Hongdong Li, Tom Drummond, and Zongyuan Ge. Hierarchical neural architecture search for deep stereo matching. In *Advances in Neural Information Processing Systems*, pages 22158–22169, 2020. 3

[6] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2758–2766, 2015. 2

[8] S. Duggal, S. Wang, W. Ma, R. Hu, and R. Urtasun. Deeppruner: Learning efficient stereo matching via differentiable patchmatch. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4383–4392, 2019. 7

[9] Y. Feng, Z. Liang, and H. Liu. Efficient deep learning for stereo matching with larger image patches. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017. 2

[10] A. Geiger, P. Lenz, and R. 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, 2012. 2, 5

[11] S. Gidaris and N. Komodakis. Detect, replace, refine: Deep structured prediction for pixel wise labeling. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7187–7196, 2017. 3

[12] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2492–2501, 2020. 7

[13] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li. Group-wise correlation stereo network. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3268–3277, 2019. 1, 3, 7, 8

[14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015. 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1, 3

[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 3

[17] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1655, 2017. 1, 4, 5

[18] Z. Jie, P. Wang, Y. Ling, B. Zhao, Y. Wei, J. Feng, and W. Liu. Left-right comparative recurrent model for stereo matching. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3838–3846. 3

[19] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry. End-to-end learning of geometry and context for deep stereo regression. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 66–75, 2017. 1, 3, 7, 8

[20] Sameh Khamis, Sean Fanello, Christoph Rhemann, Adarsh Kowdle, Julien Valentin, and Shahram Izadi. Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3, 7, 8

[21] Z. Liang, Y. Feng, Y. Guo, H. Liu, W. Chen, L. Qiao, L. Zhou, and J. Zhang. Learning for disparity estimation through feature constancy. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2811–2820, 2018. 1, 2, 3

[22] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4040–4048, 2016. 1, 2, 3, 4, 5, 7, 8

[23] Michael D Menz and Ralph D Freeman. Stereoscopic depth processing in the visual cortex: a coarse-to-fine mechanism. In *2003 Nature neuroscience*, pages 59–65, 2003. 5

[24] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3061–3070, 2015. 2, 5

[25] G. Nie, M. Cheng, Y. Liu, Z. Liang, D. Fan, Y. Liu, and Y. Wang. Multi-level context ultra-aggregation for stereo matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3278–3286, 2019. 3

[26] J. Pang, W. Sun, J. S. Ren, C. Yang, and Q. Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *2017 IEEE International Conference*

on Computer Vision Workshops (ICCVW), pages 878–886, 2017. 1, 3, 4, 5

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8026–8037. Curran Associates, Inc., 2019. 5

[28] O. Ronneberger, P. Fischer, and T. Brox. U-net:convolutional networks for biomedical image segmentation. In *2015 Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 3

[29] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesic, and P. Westling X Wang. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition (GCPR)*, 2014. 6

[30] D. Scharstein, R. Szeliski, and R. Zabih. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, 2001. 2

[31] X. Song, X. Zhao, H. Hu, and L. Fang. Edgestereo: A context integrated residual pyramid network for stereo matching. In *2019 Asian Conference on Computer Vision (ACCV)*, 2018. 1, 2, 3

[32] C. Stucker and K. Schindler. Resdepth: Learned residual stereo reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 707–716, 2020. 3

[33] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano. Real-time self-adaptive deep stereo. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 195–204, 2019. 1, 7

[34] Qiang Wang, Shaohuai Shi, Shizhen Zheng, Kaiyong Zhao, and Xiaowen Chu. FADNet: A fast and accurate network for disparity estimation. In *2020 IEEE International Conference on Robotics and Automation (ICRA 2020)*, pages 101–107, 2020. 1, 2, 3, 4, 5, 7, 8

[35] Q. Wang, S. Zheng, Q. Yan, F. Deng, K. Zhao, and X. Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021. 2

[36] Y. Wang, Z. Lai, G. Huang, B. H. Wang, L. van der Maaten, M. Campbell, and K. Q. Weinberger. Anytime stereo image depth estimation on mobile devices. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5893–5900, 2019. 3

[37] H. Xu and J. Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1956–1965, 2020. 1, 3, 5, 7, 8

[38] Guorun Yang, Hengshuang Zhao, Jianping Shi, Zhidong Deng, and Jiaya Jia. Segstereo: Exploiting semantic infor-

mation for disparity estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 7, 8

[39] Z. Yin, T. Darrell, and F. Yu. Hierarchical discrete distribution decomposition for match density estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6037–6046, 2019. 7

[40] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1983–1992, 2018. 1

[41] Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(65):1–32, 2016. 2

[42] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 7, 8

[43] J. Žbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1592–1599, 2015. 7, 8