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

Songyan Zhang¹, Zhicheng Wang¹, Qiang Wang², Jinshuo Zhang¹, Gang Wei¹, Xiaowen Chu²
¹Tongji University, ²Hong Kong Baptist University

Abstract

Existing state-of-the-art disparity estimation works mostly leverage the 4D concatenation volume and construct a very deep 3D convolution neural network for disparity regression, which is inefficient considering the high memory consumption and slow inference speed. In this paper, we propose a network named EDNet for efficient disparity estimation. To be specific, we construct a combination volume which incorporates contextual information from the concatenation volume and feature similarity measurement from the correlation volume. The combination volume can be aggregated by 2D convolutions which require less running memory. We further propose a spatial attention based residual learning module to generate attention-aware residual features. Accurate disparity correction can be provided even in low-texture regions as the residual learning process can specifically concentrate on inaccurate regions. Extensive experiments on Scene Flow and KITTI datasets show that our network outperforms previous 3D convolution based works and achieves state-of-the-art performance with significantly faster speed and less memory consumption, demonstrating the effectiveness of our proposed method.

1. Introduction

Accurate depth is of great significance to many applications like robot navigation, 3D reconstruction and autonomous driving. The target of stereo matching is to obtain the correspondence between pixels from stereo images and compute the disparity \( d \) for each pixel. Depth can be then calculated by \( (fB/d) \), where \( f \) is the camera’s focal length and \( B \) is the distance between two camera centers.

There exist two popular methods using convolution neural network (CNN) to represent and aggregate features for further disparity regression. DispNetC [2] directly regresses disparity by constructing a correlation volume from left and right features following a 2D CNN. GCNet [11] is the first to build a concatenation feature volume and utilize 3D CNN to aggregate contextual features. However, as Guo et al. point out in [13], the full correlation method produces a correlation map with only a single channel for each disparity level and thus loses much information even though it provides an efficient way for measuring feature similarities. The concatenation based method can preserve more contextual information and 3D convolutions have stronger regularization ability but it has to learn the similarity measurement function from scratch, which is inefficient. Moreover, the 3D aggregation network requires more parameters and running memory. According to the KITTI benchmark, it takes PSMNet [12] 0.41s to predict a KITTI stereo pair and GANet-deep [28] costs 1.8s to obtain top-performing results. By contrast, DispNetC [2] only needs 0.06s for prediction which is considerably faster. How to achieve the complementary advantages of the concatenation method and correlation method in an efficient way remains unsolved.

Since the residual learning strategy can simplify the training process [17], many recent works internally split the computation into an initial, coarse disparity estimation and a subsequent residual refinement [4][7][30]. Learning an additive residual which is mostly small is easier than learning the entire disparity process. But the drawback is that the conventional residual learning method has no explicit evidence about where the stereo matching algorithm is failed. We argue that the it could be more efficient if a spatial guidance of error is provided.
In order to address the above issues, we propose a combination volume and spatial attention based residual learning module to improve the efficiency for stereo matching. The proposed combination volume alleviates the information loss by employing the aggregated concatenation volume and preserves the feature similarity measurement as the correlation volume is utilized. We adopt 2D convolutions for further aggregation so that additional memory consumption and computational complexity can be avoided. Inspired by the attention mechanism, we adopt a spatial attention module to generate attention-aware residual features. Therefore, the residual learning module can have an intuitive guidance about inaccurate regions so that a specific correction can be computed, which improves the learning efficiency. Our network can generate accurate and continuous disparity map even in low-texture regions. Figure 1 visualizes the residual learning process. Our network produces more accurate prediction than other top-performing works.

The contributions of our work can be summarized as follows:

- We propose a combination volume to preserve contextual information and feature similarities, which can be aggregated with efficient 2D convolutions.
- We perform residual learning with spatial attention mechanism and provide an intuitive spatial guidance to improve the learning efficiency.
- Compared with 3D based methods, our network achieves state-of-the-art performance on Scene Flow [2] and KITTI datasets [35][40] with less memory consumption and faster inference speed.

2. Related Works

Popular stereo matching pipeline consists of four steps [1]. Over the last few years, CNN has drawn great attention and been introduced to tackle problems in stereo matching [22]. In this section, we will first briefly discuss two common mechanisms for computing the matching cost with CNN and then review the approaches with residual learning method.

2.1. Matching Cost Computation

CNN based matching cost computation methods make great contribution to the stereo matching accuracy. There are two popular methods for matching cost computation. The first one is using either a layer of 2D [3] or 1D [2] convolutional operations, called correlation layer. Such inner product between feature vectors are adopted in [4][5][9][10]. Liang et al. [4] build a correlation volume for initial disparity estimation which follows a disparity refinement module by learning through feature consistency. Wang et al. [8] 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 corresponding features from the opposite stereo image across each disparity level, which follows 3D convolutions to aggregate features and regress disparity. This method can be found first in [11]. Chang et al. [12] improve Kendall’s approach [11] by designing a spatial pyramid pooling module [31] 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 overcome the drawbacks of each cost volume. The best performance on Scene Flow dataset comes from [15] which introduces the idea of DenseNet [24] to further improve PSMNet [12]. Zhang et al. propose GANet [28].
Figure 2: An overview of our proposed EDNet. We exploit the architecture of DispNetC as the backbone. The attention-based residual learning module and combination volume are proposed for accuracy improvement. In order to provide better visualization, the skip connections between encoder and decoder in DispNetC and some other data flow are omitted here.

by building two guided aggregation layers and apply fifteen 3D convolutions to achieves state-of-the-art performance.

2.2. Residual for Stereo Matching

The residual learning concept is proposed by He et al. [17] which turns to be an efficient way to train a CNN model and has been adopted by many works. In stereo matching task, the residual learning strategy is widely used for refining disparity estimation [4][8][14][30]. Pang et al. [7] 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. [18] specially build a U-Net [20] based network to enhance the reconstruction by regressing a residual correction. In order to meet the need of real-time inference, [19] takes residual learning strategy to flexibly output disparity estimation according to the requirement of applications. Song et al. [23] manage to aggregate edge information for residual learning and thus construct a multi-task network for edge detection and stereo matching.

3. Method

3.1 Network Architecture

The architecture of our proposed EDNet is shown in Figure 2. We exploit the structure of DispNetC [2] as backbone with extensive modifications. For feature extraction, the last left and right feature maps of conv3 from the weights-sharing encoder are used to form the combination volume which is composed of the concatenation volume and correlation volume. The concatenation volume is the same as PSMNet [12] but is compressed into 1 channel with three 3D convolutions while the correlation volume has no difference from DispNetC [2]. The structure of our proposed combination volume can be found in the left bottom corner of Figure 2.

2D convolutions are then employed to aggregate the combination volume. The proposed combination volume will be described in detail in Section 3.2. In decoder part,
we follow the coarse-to-fine strategy to refine the disparity progressively. 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 [12] is adopted for residual regression but is
implemented by 2D convolutions. The attention-based
residual learning module is well illustrated in the right
bottom corner of Figure 2. Different from DispNetC
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. Combination Cost Volume

Given a pair of stereo features \(f_L\) and \(f_R\), we follow the
1D-correlation in DispNetC [2] to calculate the
correspondence at each disparity level \(d\). The correlation
volume is computed as:

\[
C_{\text{corr}}(d, x, y) = \frac{1}{N} <f_L(x - d, y), f_R(x, y)>
\]  
(1)

where \(<, >\) is the inner product of two feature vectors 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 \(H \times W\). Kendall et al. propose GCNet [11] to form a
4D feature volume by concatenating left and right features
along the disparity dimension, i.e.,

\[
C_{\text{concat}}(d, x, y, :) = \text{Concat}\{f_L(x - d, y), f_R(x, y)\}
\]  
(2)

Experimental results from [13] show that the group-wise
Correlation volume and concatenation volume are
complementary to each other. Despite the extensive
modifications for reducing the cost of 3D convolutions,
GwcNet [13] still can’t meet the need of real-time
inference. To this end, after obtaining the concatenation
volume with the shape \(N \times D \times C \times H \times W\), we use three 3D
convolutions to aggregate the concatenation volume and
compress it into 1 channel. The aggregated concatenation
volume now has the shape \(N \times D \times 1 \times H \times W\) and is then
squeezed into \(N \times D \times H \times W\), the same shape as the
correlation volume. Both the correlation volume and
squeezed concatenation volume are concatenated to
construct the final combination volume. In this way, both
the contextual information and feature similarity
measurement are incorporated into the combination
volume which follows 2D convolutions for further
aggregation instead of 3D convolutions. The final
combination volume is formed as:

\[
C_{\text{concat}}(x, y, :) = \text{Concat}\{C_{\text{corr}}(x, y), C_{\text{concat}}(x, y)\}
\]  
(3)

3.3. Attention-Based Residual Learning

Normal residual learning method lacks the spatial evidence about where the errors occur. We propose
attention-based residual learning module to guide the residual learning pay more attention to those inaccurate
regions. According to the estimated disparity \(\hat{d}\) at scale \(s\), a synthesized left image \(I^L_s\) can be obtained by warping
the right image \(I^R_s\), i.e.,

\[
I^L_s(x, y) = I^R_s(x + \hat{d}(x, y), y)
\]  
(4)

With the warped left image and target left image, we can get the error \(E^L_s = |I^L_s - I^L_s|\). A spatial attention module
with 3 layers of 2D convolution which are \(1 \times 1, 3 \times 3\) and \(1 \times 1\) respectively is applied. The spatial attention
feature map \(f^s_{\text{at}}\) is compressed into 1 channel followed by the sigmoid function to compute the spatial attention
vector whose size is \(H \times W \times 1\). We argue that with the input of color stereo image and error map, the spatial attention
module can learn an attention distribution on blurry object
boundaries and mismatched regions across the whole
spatial space. Akin to FADNet [8], CRL [7], FlowNet2 [6],
the input to spatial attention module is the concatenation of the
stereo image, error map and estimated disparity map.

To further enhance the performance, the input residual
features are concatenated with the input to spatial attention
module as shown in Figure 2. The final attention-aware
residual features \(f^s_{\text{reg}} \in R^{H \times W \times C}\) at scale \(s\) are computed by
multiplying the attention vector and residual features
\(f^s_{\text{reg}} \in R^{H \times W \times C}\), i.e.,

\[
f^s_{\text{reg}} = f^s_{\text{reg}} \cdot \sigma(f^s_{\text{at}})
\]  
(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 [12] but is implemented with 2D convolutions.

3.4. Multiscale Residual Learning

Different from the aforementioned works [7][8], we
compute the error map and attention-based residual across
multiple scales to refine the estimation progressively In
that way, the residual learning module can learn from error maps and stereo color images at the corresponding scale. The multiscale residual outputs are denoted as \( \{r^s\}_{s=0}^{S-1} \) where 0 denotes the scale of full resolution. The initial disparity is directly regressed at the coarsest scale \( S \). For the rest \( S \) scales, the estimated disparity at 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:

\[
\hat{d}^s = \hat{d}_{up}^s + r^s, 0 \leq s \leq S - 1
\]  

\[ (6) \]

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 \( s \):

\[
L^s(d^s, \hat{d}^s) = \frac{1}{N} \sum_{i=1}^{N} \text{Smooth}_{L1}(d_i^s, \hat{d}_i^s) \]  

\[ (7) \]

in which

\[
\text{Smooth}_{L1}(x) = \begin{cases} 
0.5x^2, & \text{if } |x| < 1 \\
|x| - 0.5, & \text{otherwise}
\end{cases}
\]  

\[ (8) \]

where \( N \) is the number of pixels of the disparity map, \( \hat{d}_i^s \) is the \( i \)th element of the predicted disparity \( \hat{d}^s \) and \( d^s \) represents the ground-truth disparity.

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

\[
L = \sum_{s=0}^{S} \lambda_s L^s(d^s, \hat{d}^s)
\]  

\[ (9) \]

where \( \lambda_s \) is a scalar for adjusting loss weight at scale \( s \).

4. Experiments

4.1. Datasets and Evaluation Metrics

Extensive experiments are conducted on three datasets: Scene Flow, KITTI 2012 and KITTI 2015. The Scene Flow dataset [2] consists of 39,824 pairs of synthetic stereo RGB images (35,454 pairs for training and 4,370 pairs for testing) whose full resolution is 960×540. Both KITTI 2012 [40] and KITTI 2015 [35] are datasets of the real world with a full resolution of 1242×375. The ground truth of these two datasets are generated by lidar so that only sparse ground truth is available. We evaluate our model on the 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 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

Our approach is implemented in PyTorch [25] and Adam [26] (momentum=0.9, beta=0.999) is used as optimizer in terms of accuracy. For the 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 [30], where the initial learning rate is set to be 0.001 and decreased by half every 10 epochs after 20th epoch. The loss weights are set to be \( \lambda_0 = 1.0, \lambda_1 = \lambda_2 = 0.8, \lambda_3 = 0.6 \). The crop size for KITTI datasets is set as 256×512. As there are only 200 pairs of stereo images for KITTI 2015 and 196 pairs for KITTI 2012, the training data is insufficient. 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 to get the submission result respectively. We use a constant learning rate of 0.0001 for KITTI datasets. Inspired by [34] that searching the correspondence at a coarse scale can be beneficial, especially in low-texture or textureless regions. The scalars for loss weight 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 [27] for data pre-processing. The maximum disparity is set as 192 pixels.

4.3. Ablation Study

We conduct a thorough analysis of our proposed EDNet with different settings. All the experiment results are obtained by training 10 epochs on the Scene Flow dataset.

As shown in Table 1, removing either of our proposed module leads to a clear drop of performance. Figure 3 indicates that better details like sharper object boundaries can be recovered with the help of the attention-based residual learning module. Models without the concatenat-
| Method       | Cost Volume | Residual Module | Results |
|--------------|-------------|-----------------|---------|
|              | correlation | normal          | EPE     | >1px(%) | >3px(%) |
| EDNet-NRC    | ✓           |                 | 1.67    | 27.9    | 9.7     |
| EDNet-NR     | ✓           | ✓               | 1.63    | 27.4    | 9.8     |
| EDNet-NA     | ✓           | ✓               | 1.04    | 12.7    | 5.4     |
| EDNet-NC     | ✓           | ✓               | 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 the Scene Flow dataset. The ablation study results show that our proposed method efficiently improves the prediction performance.

In this subsection, we compare our method with some of the existing state-of-the-art methods from the aspects of runtime, memory consumption and accuracy on the Scene Flow, KITTI 2015 and KITTI 2012 datasets. We further test the generalization abilities on the Middlebury 2014 dataset [41].

The inference time is tested on the Scene Flow dataset with a resolution of 576×960. As can be analyzed from Table 2, our method has the best score among all competing methods. Our proposed EDNet not only outperforms some of the state-of-the-art methods in accuracy which is refle-
Table 2: Several state-of-the-art methods EPE evaluation results on Scene Flow. Our method achieves the state-of-the-art performance. All the running times are measured on a single NVIDIA 2080ti GPU for fair comparison. OOM denotes out of memory.

| Method       | PSMNet[12] | GANet[28] | GwcNet[13] | Bi3D[36] | StereoNet[14] | DispNetC[2] | AAANet+[30] | Ours       |
|--------------|------------|-----------|------------|----------|---------------|-------------|-------------|------------|
| EPE          | 1.68       | 0.84      | 0.76       | 0.73     | 1.10          | 1.68        | 0.72        | **0.63**   |
| Time(s)      | 0.453      | 3.302     | 0.254      | OOM      | 0.034         | 0.025       | 0.068       | **0.059**  |

Table 3: Benchmark results on KITTI2015 test sets. “Noc” and “All” indicates percentage of outliers averaged over ground truth pixels of non-occluded and all regions respectively. “fg” and “all” indicate percentage of outliers averaged over foreground and all ground truth pixels respectively.

| Method                | Noc(%) | All(%) | Time(s) |
|-----------------------|--------|--------|---------|
| GANet[28]             | 3.37   | 1.73   | 3.82    | 1.93    | 0.36          |
| GCNet[11]             | 5.58   | 2.61   | 6.16    | 2.87    | 0.9           |
| PSMNet[12]            | 4.31   | 2.14   | 4.62    | 2.32    | 0.41          |
| GwcNet[13]            | 3.49   | 1.92   | 3.93    | 2.11    | 0.32          |
| SegStereo[5]          | 3.70   | 2.08   | 4.07    | 2.25    | 0.6           |
| MC-CNN[29]            | 7.64   | 3.33   | 8.88    | 3.89    | 0.7           |
| HD3[32]               | 3.43   | 1.87   | 3.63    | 2.02    | 0.14          |
| ActNet[37]            | 3.49   | 1.72   | 3.80    | 1.51    | 0.48          |
| CSN[38]               | 3.55   | 1.78   | 4.03    | 1.59    | 0.5           |
| DeepPruner-B[39]      | 3.18   | 1.95   | 3.56    | 2.15    | 0.18          |
| Bi3D[36]              | 3.11   | 1.79   | 3.48    | 1.95    | 0.48          |
| Ours                  | **3.33**| 2.31   | 3.88    | 2.53    | **0.05**      |
| AANet[30]             | 4.93   | 2.32   | 5.39    | 2.55    | 0.075         |
| DeepPruner-F[39]      | 3.43   | 2.35   | 3.91    | 2.59    | 0.06          |
| StereoNet[14]         | \     | \     | 7.45    | 4.83    | **0.015**     |
| DispNetC[2]           | 3.72   | 4.05   | 4.41    | 4.34    | 0.06          |
| FADNet[8]             | **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 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               | Out(%) | Avg | Time(s) |
|----------------------|--------|-----|---------|
| GANet[28]            | 2.18   | 0.5 | 0.36    |
| GCNet[11]            | 2.71   | 0.6 | 0.7     |
| PSMNet[12]           | 2.44   | 0.5 | 0.6     |
| GwcNet[13]           | 2.16   | 0.5 | 0.5     |
| SegStereo[5]         | 2.66   | 0.5 | 0.6     |
| MC-CNN[29]           | 3.90   | 0.8 | 1.0     |
| ActNet[37]           | **1.83**| 0.5 | 0.5     |
| Ours                 | 2.97   | 0.5 | 0.6     |
| AANet[30]            | **2.90**| 0.5 | 0.6     |
| StereoNet[14]        | 4.91   | 0.8 | 0.9     |
| DispNetC[2]          | 7.36   | 0.9 | 1.0     |
| FADNet[8]            | 3.98   | 0.6 | 0.7     |
| Ours                 | 2.97   | 0.5 | 0.6     |

Our method improves the performance by 60% and 40% respectively. Figure 4 gives a visual comparison between our EDNet and other top-performing methods on the Scene Flow dataset. Our proposed method generates sharper object boundaries and more continuous disparity map, indicating the value of our proposed approach.

As for the KITTI 2015 and KITTI 2012 datasets, we div-
the-art works.

Figure 5: Results of 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 [2], PSMNet [12], GwcNet [13], AANet[30] and our EDNet, as well as their error maps.

| Method     | Memory | FLOPs   | Time(s) |
|------------|--------|---------|---------|
| PSMNet[12] | 4.83G  | 937.90G | 0.393   |
| GANet[28]  | 6.53G  | 1936.98G| 2.430   |
| GwcNet[13] | 4.27G  | 899.99G | 0.272   |
| Bi3D[36]   | 10.74G | 4212.05G| 0.899   |
| Ours       | 2.52G  | 162.92G | 0.053   |

Table 5: Comparisons of the runtime, running memory and computational cost. All the results are tested on a NVIDIA RTX 2080Ti GPU with popular 3D convolution based models at a resolution of 1248 × 384.

Table: Generalization on Middlebury 2014 dataset. The bounding boxes emphasize the effectiveness of our proposed method.

| Method     | Memory | FLOPs   | Time(s) |
|------------|--------|---------|---------|
| PSNMNet[12]| 4.83G  | 937.90G | 0.393   |
| Bi3D[36]   | 10.74G | 4212.05G| 0.899   |
| Ours       | 2.52G  | 162.92G | 0.053   |

Figure 6: Generalization on Middlebury 2014 dataset. The bounding boxes emphasize the effectiveness of our proposed method.

5. Conclusion

We have exploited an efficient architecture called EDNet with the proposed combination volume and attention-residual learning module. We show that the combination of the contextual information and feature similarity measurement is of great importance to robust feature representations, especially in ill-posed regions. Besides, the learning of residual can be more efficient with a spatial attention guidance. Extensive experimental results on the 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 improving our method with a more lightweight design and applying our work to other depth-related tasks, e.g., 3D reconstruction, robot navigation.

The Middlebury 2014 dataset is utilized for
Reference

[1]. D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” IJCV, vol. 47, no. 1-3, pp. 7–42, 2002.

[2]. N. Mayer, E. Ilg, P. Hausser, 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 The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 1, 2, 3

[3]. A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. “FlowNet: Learning optical flow with convolutional networks,” in IEEE ICCV, 2015.

[4]. Liang, Z., Feng, Y., Guo, Y., Liu, H., Chen, W., Qiao, L., Z., L., Z., J.: Learning for disparity estimation through feature constancy. In: CVPR (2018). 1 2

[5]. G. Yang, H. Zhao, J. Shi, Z. Deng, and J. Jia. Segstereo: Exploiting semantic information for disparity estimation. In Proceedings of the European Conference on Computer Vision (ECCV 2018).

[6]. 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 Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2462–2470, 2017.

[7]. 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 ICCV Workshops, volume 7, 2017. 1 2

[8]. Q. Wang, S. Shi, S. Zheng, K. Zhao, X. Chu. Fadnet: A Fast and Accurate Network for Disparity Estimation. In ICRA 2020.

[9]. W. Luo, A. G. Schwing, and R. Urtasun, “Efficient deep learning for stereo matching,” in IEEE CVPR, 2016.

[10]. W. Chen, X. Sun, L. Wang, Y. Yu, and C. Huang, “A deep visual correspondence embedding model for stereo matching costs,” in IEEE ICCV, 2015.

[11]. 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 Proceedings of the IEEE International Conference on Computer Vision, pages 66–75, 2017.

[12]. 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.

[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.

[14]. S. Khamis, S. Fanello, C. Rhemann, A. Kowdle, J. Valentin, and S. Izadi, “Stereonet: Guided hierarchical refinement for real-time edge-aware depth prediction," ECCV, 2018.

[15]. Guang-Yu Nie, Ming-Ming Cheng, Yun Liu, Zhengfa Liang, Deng-Ping Fan, Yue Liu, and Yongtian Wang. Multi-level context ultra-aggregation for stereo matching. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3283–3291, 2019.

[16]. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., volume 1, page 3, 2017.

[17]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[18]. C. Stucker, K. Schindler.ResDepth: Learned Residual Stereo Reconstruction. arXiv:2001.08026v1 [cs.CV] 22 Jan 2020.

[19]. Y.Wang, Z.Lai, G.Huang, B.H, Anytime Stereo Image Depth Estimation on Mobile Devices.arXiv:1810.11408v2 [cs.CV] 5 Mar 2019.

[20]. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proceedings of the IEEE Conference on Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.

[21]. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. MICCAI, 2015.

[22]. J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 17(1-32):2, 2016.

[23]. Xiao Song, Xu Zhao, Hanwen Hu, and Liangji Fang. EdgeStereo: A context integrated residual pyramid network for stereo matching. Proceedings of the Asian Conference on Computer Vision (ACCV), 2018.

[24]. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In IEEE Conf. Comput. Vis. Pattern Recog., volume 1, page 3, 2017.

[25]. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems, pages 8024–8035, 2019.

[26]. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[27]. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 248–255.

[28]. Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip H.S. Torr. GA-Net: Guided aggregation net for end-to-end stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[29]. 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 (CVPR), 2015.

[30]. Haofei Xu, Juyong Zhang. AANet: Adaptive Aggregation Network for Efficient Stereo Matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[31]. K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell., 37(9):1904–1916, 2015.

[32]. Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6044–6053, 2019.

[33]. Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 195–204, 2019.

[34]. Michael D Menz and Ralph D Freeman. Stereoscopic depth
processing in the visual cortex: a coarse-to-fine mechanism. *Nature neuroscience*, 6(1):59–65, 2003.

[35]. M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.

[36]. A. Badki, A. Troccoli, K. Kim, J. Kautz, P. Sen and O. Gallo: Bi3D: Stereo Depth Estimation via Binary Classifications. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2020.

[37]. Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li and K. Yang: Adaptive Unimodal Cost Volume Filtering for Deep Stereo Matching. AAAI 2020.

[38]. 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2020.

[39]. Shivam Duggal, Shenlong Wang, Wei-Chi 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.

[40]. 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.

[41]. Daniel Scharstein, Heiko Hirschmuller, York Kitajima, Greg Krathwohl, Nera Nesić, 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.