Abstract-- In this paper, we propose novel networks for stereo disparity estimation. First, deep features are extracted using efficient depthwise-separable convolutions. Next, the stereo matching costs are calculated from the deep features with a novel extended cost volume. Then, rich multiscale contextual information is aggregated with the proposed atrous multiscale network (AMNet). The proposed foreground-background aware network (FBA-AMNET) is trained with an iterative multi-task learning strategy to discriminate between foreground and background objects at multiple scales. The proposed networks advance the state of the art on challenging disparity estimation benchmarks, such as the KITTI 2012, KITTI 2015, Sceneflow, and Middlebury stereo benchmarks.

I. INTRODUCTION

Stereo disparity estimation has many applications such as depth estimation for scene understanding and the synthesis of the Bokeh effect in images captured with dual camera phones. Often the stereo images are first rectified to lie in the same image plane, such that corresponding pixels in the left and right images lie on the same horizontal line. Disparity estimation pipelines classically consist of three or four steps; feature extraction, matching-cost computation, disparity aggregation and computation, and disparity refinement [1]. Recent deep learning based approaches [2,3] have achieved the state-of-art performances on public stereo datasets [4,5,6].

II. ATROUS MULTISCALE NETWORKS

In this work, we propose a novel deep neural network architecture that improves these different steps of stereo disparity estimation. The overall network architecture is shown in Fig. 2.

A. D-ResNet Feature Extractor

First, we design a depthwise separable residual network (D-ResNet) as the feature extractor that has a large receptive field, while maintaining the low complexity. The feature extractor is designed by modifying the residual blocks of the residual network of 50 layers (ResNet-50) to be formed of depthwise-separable convolutions (DW-SC). In DW-SC [8], the 3D convolutions are replaced by 2D convolutions for each channel followed by a combination layer. The low complexity of DW-SC allows us to increase the dimensions of the feature maps at the output of a residual block without increasing the complexity. Hence, a pointwise projection layer is added in the shortcut connection of each depthwise residual block to match its input and output dimensions.

B. Atrous Multiscale Context Aggregation

Second, we propose a stacked atrous multiscale (AM) module to capture deep local features, as well as global contextual information from the feature extractor at multiple scales. Depth or disparity estimation networks tend to use down-samplings and up-samplings or encoder-decoder architectures, also called hour-glass architectures [2,4] to aggregate information at multiple scales. However, the spatial resolution tends to be lost by pooling or down-sampling. Thus, we design an AM module as a set of 3x3 atrous convolutions with increasing dilation factors such as \([1,2,4,8,16,32]k/2\). The dilation factors increase as the AM module goes deeper to increase the receptive field and capture denser multiscale contextual information without losing the spatial resolution. Several of these modules can be stacked together as shown in Fig. 1.

C. Extended Cost Volume for Matching-Cost Computation

Third, the matching-cost computation is done with a novel extended cost volume (ECV) that carries rich information about the matching costs using different similarity measures. The proposed ECV constitutes of several sub-volumes. A disparity-shifted feature concatenation sub-volume concatenates the left feature maps \(F_l(d)\) and the disparity-aligned right feature map \(F_r(d)\) at each disparity \(d \in \{0,1,\ldots,D\}\), where \(D\) is maximum considered disparity, and has size of \(2(D + 1)CWH\) when the width, height and number of channels of the left feature maps are \(W,H,C\) respectively. A disparity-level depthwise correlation sub-volume computes the patch-wise correlation between the left feature maps and the corresponding right feature maps after disparity alignment at each \(d\), and thus has size of \((D + 1)CWH\). The disparity-level feature distance sub-volume computes \(D + 1\) distance maps between each left and right feature map after disparity alignment, and has size \((D + 1)CWH\). The three subvolumes are packed together into the ECV, which has a size of \(4(D + 1)CWH\). The output of the ECV is then aggregated at multiple scales by the stacked AM module shown in Fig. 1.

Fig. 1. Architecture of a stacked atrous multiscale module
D. Disparity Computation and Optimization

Disparity optimization is done by regression of the soft output obtained after the classification of the disparity into one of the quantized disparity bins. To enhance the cost volume filtering, and improve the disparity computation and optimization steps, we propose the foreground-background aware AMNet (FBA-AMNet). We devise an iterative training strategy using multi-task learning, in which the main task is disparity estimation and the auxiliary task is foreground-background segmentation. The learned foreground background information reinforces the disparity estimation.

We propose an iterative method to train FBA-AMNet. After each epoch, the latest estimated segmentation maps are concatenated with the RGB input to form an RGB-S input to the FBA-AMNet at the next epoch. During training, the network keeps refining and utilizing its foreground-background segmentation predictions, so as to learn better awareness of foreground objects. At the inference stage, to save computations, zero maps can be used as the segmentation input map. However, the learnt foreground-background awareness improves the accuracy of disparity estimation, especially around object boundaries, as shown in in Fig. 3.

III. EXPERIMENTAL RESULTS AND CONCLUSIONS

The proposed AMNets ranked first, at time of submission, among all published results on the three most popular disparity estimation benchmarks: KITTI stereo 2015 [6], KITTI stereo 2012 [5], and SceneFlow [4] stereo disparity tests. We also achieved competitive results on the Middlebury dataset [7].

FBA-AMNet’s foreground (FG) and background (BG) disparity error percentages, as published on the KITTI 2015 Stereo leaderboard is shown in Table 1. Our AMNet can also achieve state of art performance on the SceneFlow dataset with an end-point-error (EPE) of 0.74 which is 32.1% better than the best published results. Our novel networks can also achieve 106 on the A99 dense test metric on Middlebury, which ranks it first among all submissions using quarter resolution.

<table>
<thead>
<tr>
<th>METHOD</th>
<th>D1- BG%</th>
<th>D1-FG%</th>
<th>D1- ALL%</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBA-AMNET</td>
<td>1.53</td>
<td>3.43</td>
<td>1.84</td>
</tr>
<tr>
<td>AMNET</td>
<td>1.60</td>
<td>3.81</td>
<td>1.97</td>
</tr>
<tr>
<td>SEGSTEREO [3]</td>
<td>1.88</td>
<td>4.07</td>
<td>2.25</td>
</tr>
<tr>
<td>PSMNET [2]</td>
<td>1.86</td>
<td>4.62</td>
<td>2.32</td>
</tr>
</tbody>
</table>

REFERENCES