Supplementary Appendix for ”Improved Stereo Matching with Constant Highway Networks and Reflective Confidence Learning”

Amit Shaked[1] and Lior Wolf[1,2]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Israel
[2]Facebook AI Research

A. The benefit of color

Previous architectures in the literature for computing the matching cost report no benefit from using color information [3, 1]. In our experiments we observed that after deepening our network, the use of the three input channels contributes to the accuracy of the disparity prediction, especially around areas of delicate color differences between the object and its background. An example for this phenomenon is shown in Fig. 1, and the average improvements over the validation sets of KITTI and Middlebury are presented in Tab. 1.

B. Multilevel constant highway connections

We study the effect of the added inner and outer constant highway connections. The amount of the input that is added to each outer block, according to Eq. 2 in the paper, is determined by $\lambda_0 + (\lambda_1 \cdot \lambda_2)$. When the network is interpreted as an ensemble of the possible paths [2], the value of $\lambda_{i,0} + (\lambda_{i,1} \cdot \lambda_{i,2})$ for outer-block i determines the contribution of the sub-network that consists of i outer-blocks to the ensemble.

Fig. 2 depicts the progression of these values by epochs for our five outer-blocks description network. One can observe that when the network is fully trained after epoch 14, the values of the deeper outer-blocks are higher than the shallow outer-blocks. This means that the low-level blocks are hardly skipped, while the information in the fully trained network tends to skip the upper blocks more. This phenomenon is increasing with the training epochs: as training progresses, the skip values of the high-level blocks are increasing and the values at the shallower layers are decreasing.

C. Reflective confidence

In the paper, for lack of space, we only provide the AUC of the confidence scores per validation image for the KITTI 2015 benchmark (Fig. 4 of the main submission file). For completeness we provide the results for both KITTI 2012 and KITTI 2015 in supplementary Fig. 3.

D. Runtime

We measure the runtime required forthe disparity map computation of an image pair, on a computer with a single NVIDIA Titan X (Pascal) graphics processor unit. The imagesare taken from the KITTI 2012 data set and their sizes are 1242×350, with 228 possible disparities.

Tab. 2 presents the measured runtime of a single computation of each step in our accurate and fast pipelines, as well as the number of iterations required for the full prediction. The three main differences between the pipelines are: (i) The fast description sub-network contains four outer blocks and the accurate five. (ii) The fast decision sub-network contains the dot product between the two image descriptors, as described in Sec. 3 of the paper. (iii) No cost aggregation is performed in the fast pipeline.

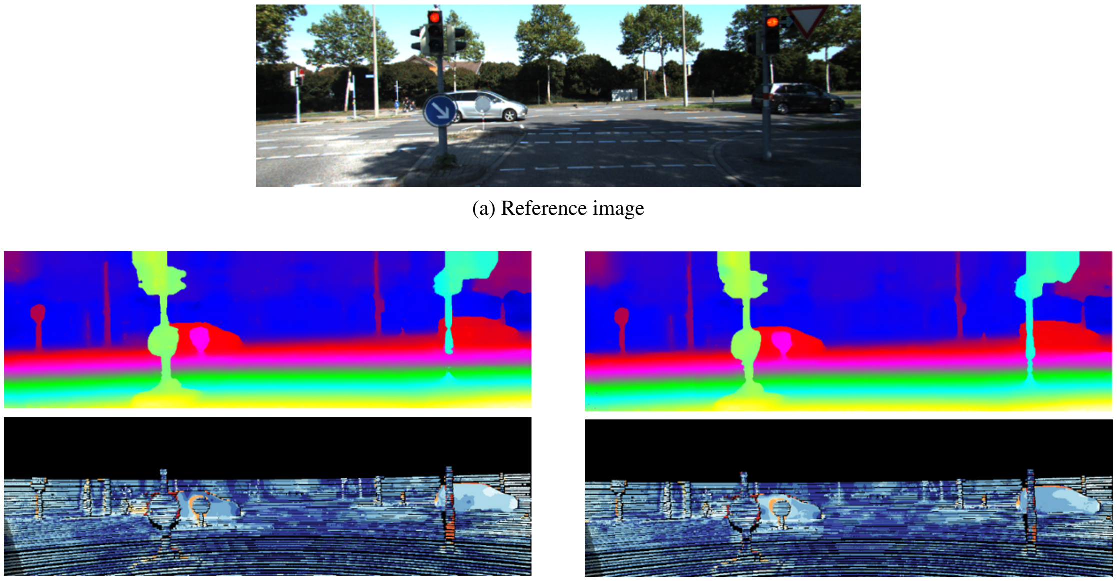
The total runtime is 48 seconds for our accurate method and 2.84 seconds for the fast. However, this can be much reduced by parallel computation of the two bottlenecks. The first is the computation time of the left and right image descriptors, which is 40 percent of the fast method's runtime, and can be reduced by computing the two descriptors in parallel. The second is the time of $disparity\_max$ forward passes in the decision sub-network, that are required in order to compute the matching cost for every possible disparity. Each forward pass can be computed in parallel, and thus reduce up to 90 percent of the accurate method's runtime and another 11 percent of the fast method's runtime.

References

[1] W. Luo, A. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

	KITTI 2012		KITTI 2015		MB
	Fast	Accurate	Fast	Accurate	Fast
mc-cnn [3]	3.02	2.61	3.99	3.25	9.87
mc-cnn [3]+color	3.02	2.61	3.99	3.25	9.87
λ-ResMatch (no color)	2.82	2.51	3.79	3.18	9.35
λ-ResMatch	2.73	2.45	3.69	3.15	9.08

Table 1: The benefits from color use in λ-ResMatch and the baseline MC-CNN. The errors reported are the validation errors after applying the post processing steps used in [3].



(a) Reference image

(b) Disparity map prediction and its errors before incorporating color information

(c) Disparity map prediction and its errors after incorporating color information

Figure 1: An example taken from KITTI 2015 data set showing the effect of color in the λ-ResMatch architecture. Observe the errors where the color differences between the traffic light and the background are very delicate.

[2] A. Veit, M. Wilber, and S. Belongie. Residual networks behave like ensembles of relatively shallow networks. In *NIPS*, 2016.

[3] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *CoRR*, abs/1510.05970, 2015.

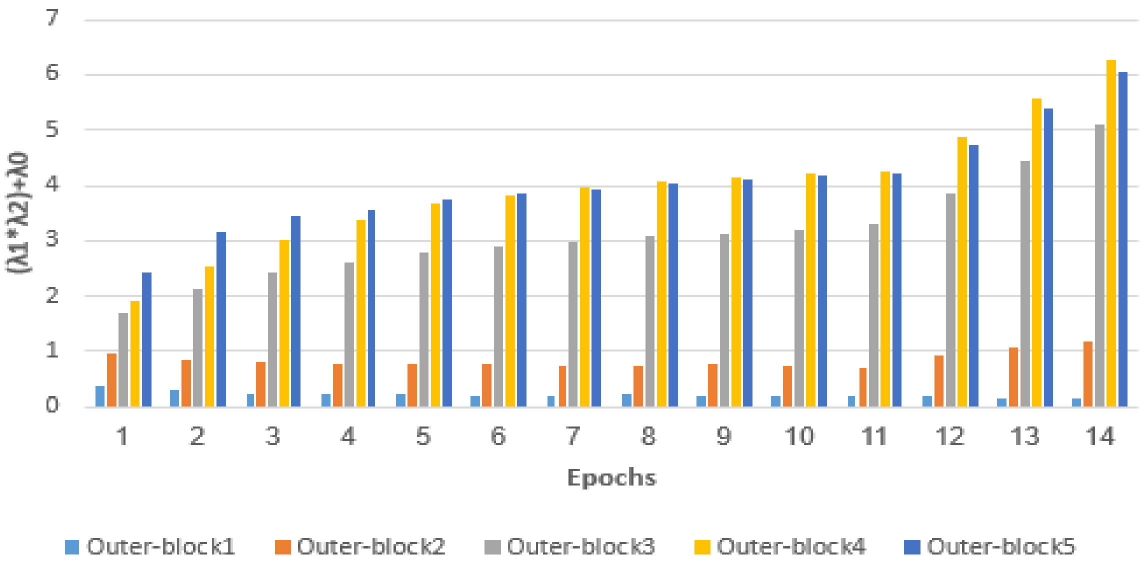
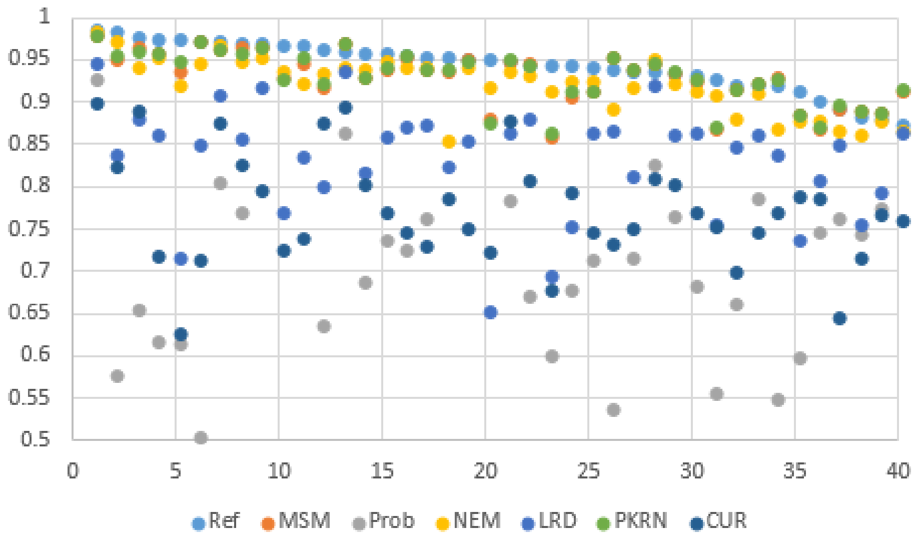


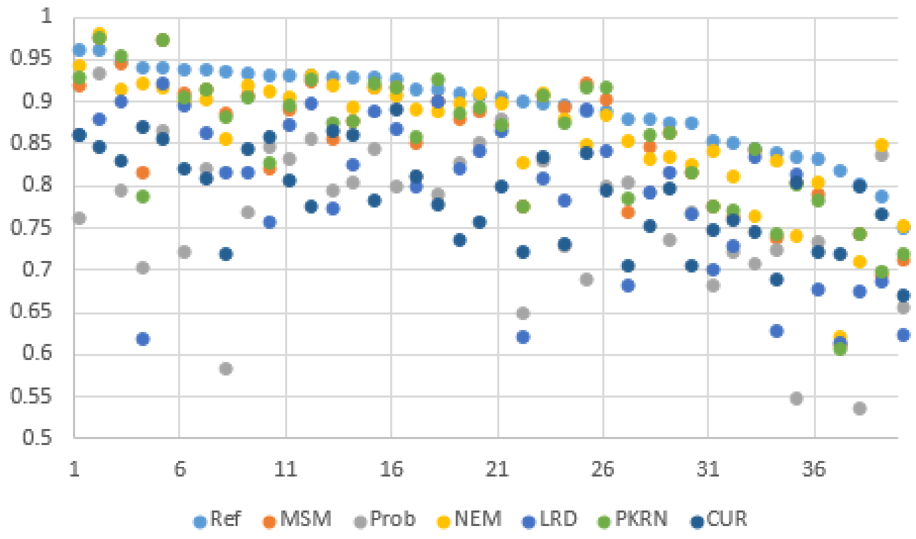
Figure 2: The values of total skip-connections for each outer clock, i.e. the amount of input that skips each outer-block through the constant highway gates, as a function of the training epochs.

	Fast			Accurate		
Component	Runtime	Iterations	Total	Runtime	Iterations	Total
Description sub-network	0.61	2	1.22	0.90	2	1.8
Decision sub-network	0.0007	456	0.31	0.097	456	44.23
CBCA	-	0	0	0.08	4	0.32
SGM	0.12	2	0.24	0.12	2	0.24
Global disparity network	1.04	1	1.04	1.04	1	1.04
Outlier interpolation	0.0008	1	0.0008	0.0008	1	0.0008
Sub-pixel enhancement	0.0001	1	0.0001	0.0001	1	0.0001
Smoothing and refinement	0.01	1	0.01	0.01	1	0.01
Everything else	0.02	-	0.02	0.02	-	0.02
Total			**2.84s**			**47.93s**

Table 2: The runtime in seconds that is required for prediction of each component on a single NVIDIA Titan X (Pascal). In order to compute the matching cost map on the KITTI data set, the description network has to be run twice: once to create the left image descriptors and once to create the right image descriptors. The decision network has to be run $disparity\_max =$ 228 times, once for every possible disparity. In order to perform the left-right consistency check in the outlier interpolation step, the matching cost map is computed twice: once when using the left image as a reference image and once when using the right image as reference. The two image descriptors can be reused, but another $disparity\_max$ forward passes in the decision network are required, followed by CBCA and SGM computations.



(a) KITTI 2012



(b) KITTI 2015

Figure 3: AUC of confidence measures on 40 random validation images from the KITTI 2015 and the KITTI 2012 stereo data sets, ordered by the reflective confidence score. The reflective confidence (Ref) is shown to outperform the different measures, described in the paper at Sec. 6.2, on almost every image.