Improved stereo matching with **Constant Highway Networks** and **Reflective Confidence**

Amit Shaked and Lior Wolf

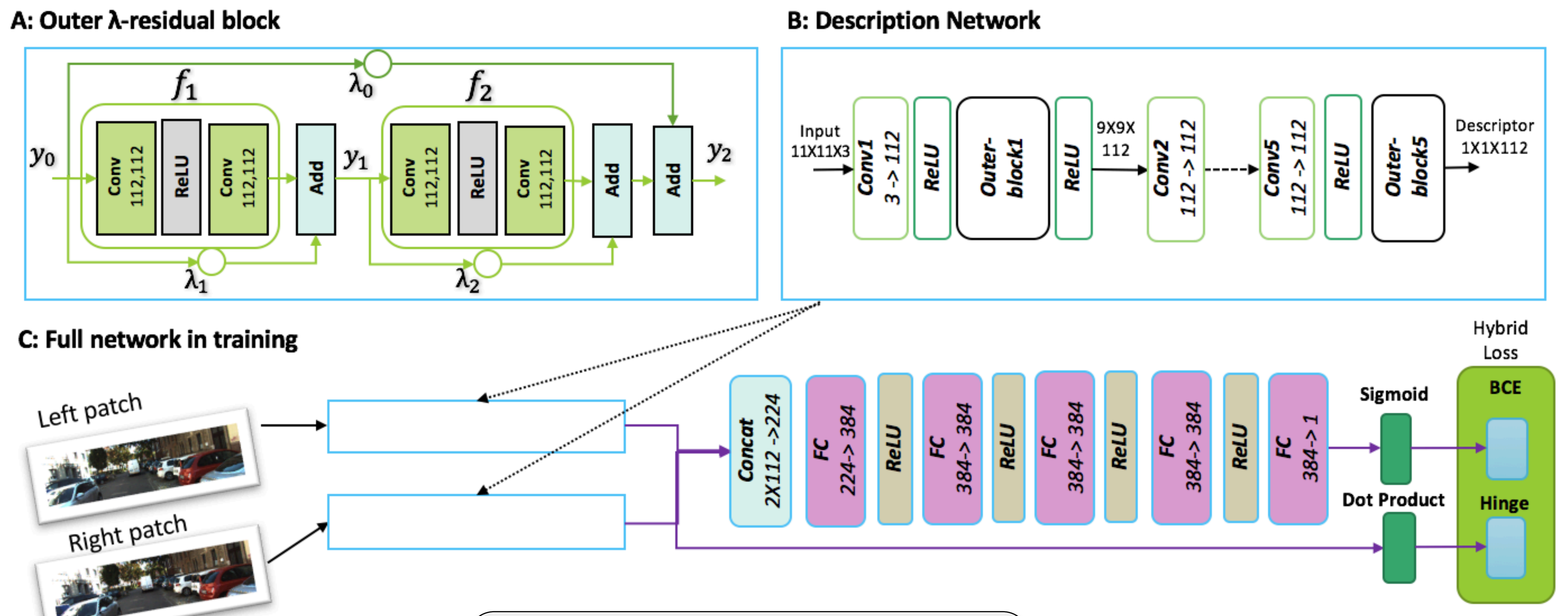
IEEE 2017 Conference on Computer Vision and Pattern Recognition

CVPR July 21-26 Honolulu 2017

Main Contributions

- **A new highway network architecture** for patch matching, suited for metric learning VS multiclass classification.
- A novel way to **measure the correctness** of the output of a CNN via **reflective learning,** that **outperforms any other technique.**
- A **CNN based post processing step** to compute the disparity image, instead of the previously suggested WTA strategy.
- A **better occlusion and mismatch detection** and interpolation.
- **Hybrid loss** for better use of **description-decision network architecture.**
- Improving the **state of the*** art in the KITTI dataset for stereo matching **by a significant margin, for both accurate and fast methods.**

Multilevel constant highway network



Control the residual block skipped data

Constant highway skip-connection:

$$y_{i+1} = f_{i+1}(y_i) + \lambda_{i+1} \cdot y_i$$

Outer λ-residual block:

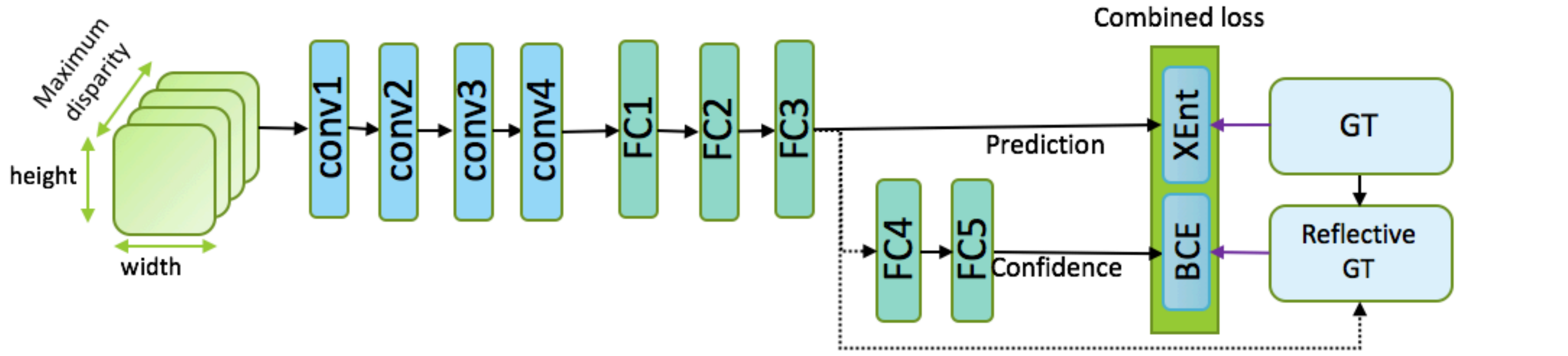
$$\begin{aligned} y_2 &= \lambda_0 y_0 + \lambda_2 \cdot y_1 + f_2(y_1) \\ &= \lambda_0 y_0 + \lambda_2(\lambda_1 y_0 + f_1(y_0)) + f_2(\lambda_1 y_0 + f_1(y_0)) \\ &= (\lambda_0 + \lambda_2\lambda_1) y_0 + \lambda_2 f_1(y_0) + f_2(\lambda_1 y_0 + f_1(y_0)) \end{aligned}$$

* At the time of writing

Reflective Confidence

GT labels change dynamically during training

Global Disparity network:



Prediction Loss**:

$$loss(\mathbf{y}, y^{GT}) = -\sum_{y_i} p(y_i, y^{GT}) \cdot \log \frac{e^{-y_i}}{\sum_j e^{y_j}}$$

$$p(y_i, y^{GT}) = \begin{cases} \lambda_1 & \text{if } |y_i - y^{GT}| \le 1 \\ \lambda_2 & \text{if } 1 < |y_i - y^{GT}| \le 2 \\ \lambda_3 & \text{if } 2 < |y_i - y^{GT}| \le 3 \\ 0 & \text{otherwise} \end{cases}$$

Reflective Loss function:

$$y_{ref}^{GT} = \begin{cases} \mathbf{1}\ \mathit{if}\ |\mathrm{argmax_i}\ \mathrm{y_i} - \mathrm{y^{GT}}| < \lambda \\ \mathbf{0}\ \mathit{otherwise} \end{cases}$$

$$loss(y_{ref}, y_{ref}^{GT}) = -(1 - y_{ref}^{GT})\ln(1 - y_{ref}) - y_{ref}^{GT} ln(y_{ref})$$

Pixel labeling:

$$\begin{array}{rl} \mathit{correct} & \text{if } \ |d - D^R(\mathbf{pd})| \le \tau_1 \ \text{ or} \\ & (C^L(\mathbf{p}) \ge \tau_2 \text{ and } C^L(\mathbf{p}) - C^R(\mathbf{pd}) \ge \tau_3) \\ \mathit{mismatch} & \text{if there exist } \hat{d} \ne d \text{ s.t. } |\hat{d} - D^R(\mathbf{p\hat{d}})| \le \tau_4 \\ \mathit{occlusion} & \text{otherwise} \end{array}$$

Where:

$C^L(\boldsymbol{p})$ - the confidence score at position $\boldsymbol{p}$ of the prediction $d = D^L(\boldsymbol{p})$

$C^L(\boldsymbol{pd})$ - the confidence score at position $\boldsymbol{p} - \boldsymbol{d}$ of the prediction $d = D^L(\boldsymbol{pd})$

Pixel Interpolation:

Mismatch - the median of the nearest neighbors labeled as correct from 16 different directions.

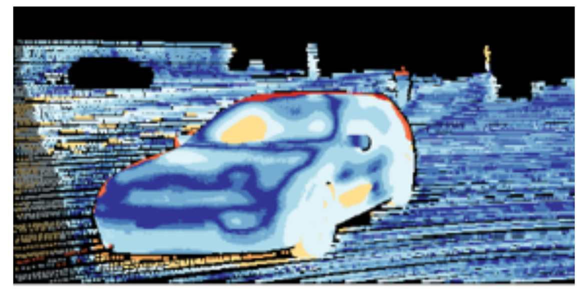
Occlusion - move left until the first correct pixel and use its value.

** The criterion is similar to [2]

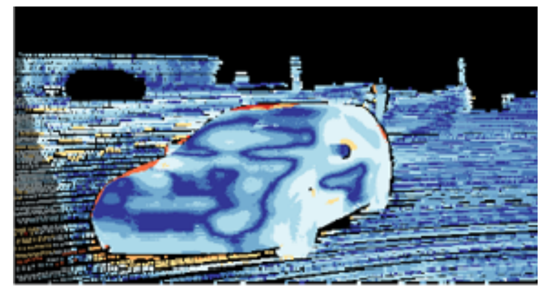
Results

A: Reference image

B: Baseline errors



C: Out solution



The highest ranking methods on KITTI:

	Method	Set.	NOC	ALL	runtime
1	**Ours**		**2.91**	**3.42**	**48s**
2	Displets v2[10]	S	3.09	3.43	265s
3	PCBP[25]		3.17	3.61	68s
4	**Ours-fast**		**3.29**	**3.78**	**2.8s**
5	MC-CNN-acrt[36]		3.33	3.89	67s

(a) KITTI 2015

	Method	Set	NOC	ALL	runtime
1	**Ours**		**2.27**	**3.40**	**48s**
2	PCBP[25]		2.36	3.45	68s
3	Displets v2[10]	S	2.37	3.09	265s
4	MC-CNN-acrt[36]		2.43	3.63	67s
5	cfusion[25]	MV	2.46	2.69	70s

(b) KITTI 2012

The highest ranking methods on KITTI for methods under 5 sec:

Rank	Method	NOC	ALL	runtime
1	**Ours-fast**	**3.29**	**3.78**	**2.8s**
2	DispNetC[22]	4.05	4.34	0.06s
3	Content-CNN[21]	4.00	4.54	1s
4	MC-CNN-fast[36]	?	4.62	0.8s
5	SGM+CNN(anon)	4.36	5.04	2s

(a) KITTI 2015

Rank	Method	NOC	ALL	runtime
1	**Ours-fast**	**2.63**	**3.68**	**2.8s**
2	MC-CNN-fast[36]	2.82	?	0.7s
3	Content-CNN[21]	3.07	4.29	0.7s
4	Deep Embed[2]	3.10	4.24	3s
5	SPS-st[34]	3.39	4.41	2s

(b) KITTI 2012

Residual architectures comparison:

	Inner shortcut	Outer shortcut	KITTI 2012	KITTI 2015	MB
mc-cnn[36]	-	-	2.84	3.53	9.73
Highway[32]	-	-	2.81	3.51	9.77
ResNet[16]	A	-	2.82	3.71	10.03
λ variant	λ	-	2.81	3.55	10.01
DC[6]	A	-	3.86	5.02	11.13
λ variant	λ	-	3.42	4.43	11.07
RoR[18]	A	C	2.86	3.52	9.68
λ variant	λ	λ· C	2.84	3.53	9.95
Variants of our method without the hybrid loss	A	A	2.78	3.49	9.63
	λ	A	2.75	3.42	9.83
	A	λ	2.78	3.46	10.3
	λ	λ	2.73	3.42	9.60
λ-ResMatch	λ	λ	**2.71**	**3.35**	**9.53**

Table 6: The validation errors of different architectures and their λ-variants, when trained on 20% of the data. "**A**" shortcut is the identity connection, "**C**" is $1X1$-convolution and "λ" is our constant highway skip-connection.

Confidence Measures Comparison:

	Ref	MSM	Prob	CUR	PKRN	NEM	LRD
KITTI2012	**0.943**	0.928	0.648	0.772	0.930	0.919	0.833
KITTI2015	**0.894**	0.850	0.758	0.832	0.853	0.864	0.812

Table 7: The average AUC over the entire validation set for different confidence measures.

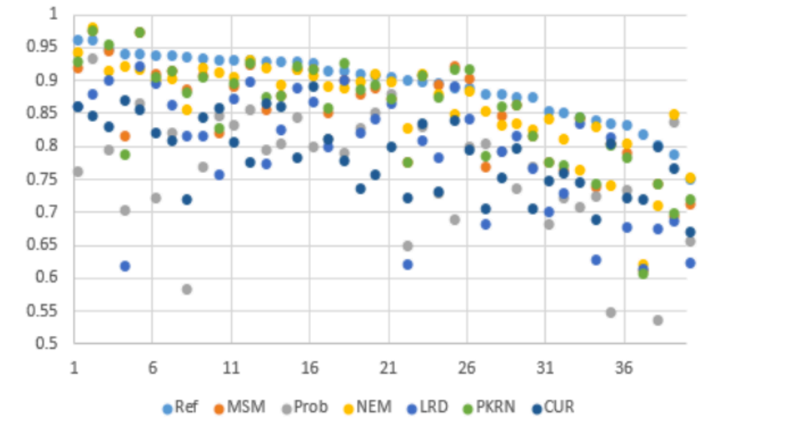


Figure 4: AUC of confidence measures on 40 random validation images from the KITTI 2015 stereo data set.

Most relevant references:

[1] J. Zbontar and Y. LeCun. **Computing the stereo matching cost with a convolutional neural network**, CVPR, 2015.

[2] W. Luo, A. Schwing, and R. Urtasun**. Efficient deep learning for stereo matching**. CVPR, 2016.

Scan for our codebase: