Efficient Deep Learning for Stereo Matching

Wenjie Luo, Alex Schwing and Raquel Urtasun

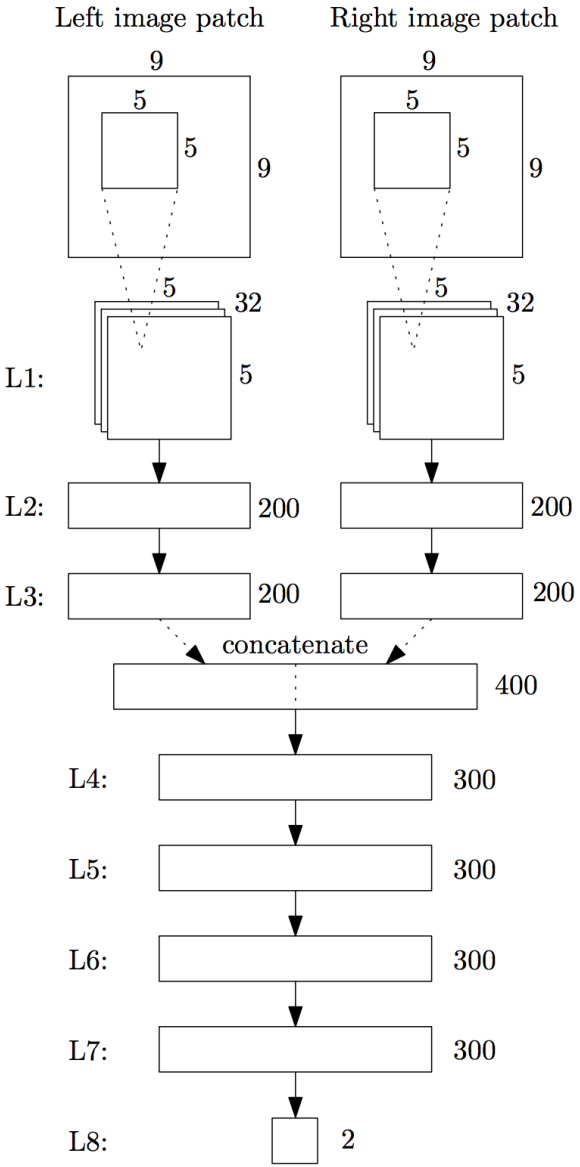
Stereo Estimation

Desired Properties:

- Good enough to detect obstacles precisely
- Fast: real time
- Robust to:
  - Saturation
  - Shadows
  - Repetitive patterns
  - Specularities
  - etc

Can we leverage deep learning to do stereo estimation?

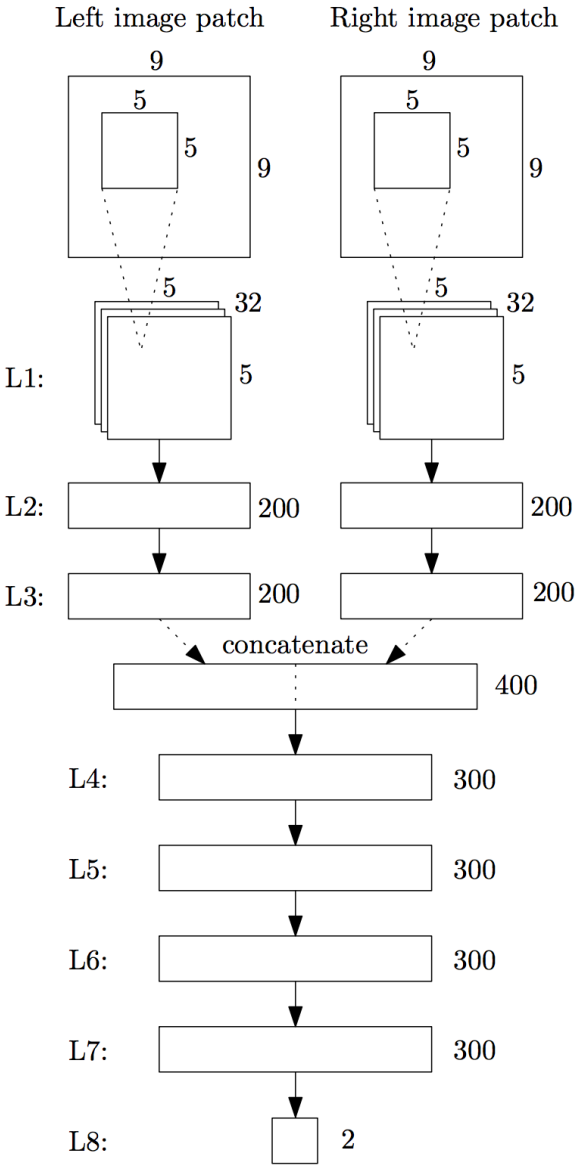
Current Deep Learning Approaches



[J. Zbontar and Y. LeCun, CVPR15]

- Current approaches use a siamese network
- Combine the two branches via concatenation follow by further processing
- Treat the problem as classification (i.e., given a left and right image patches, are they a true match?)

Current Deep Learning Approaches

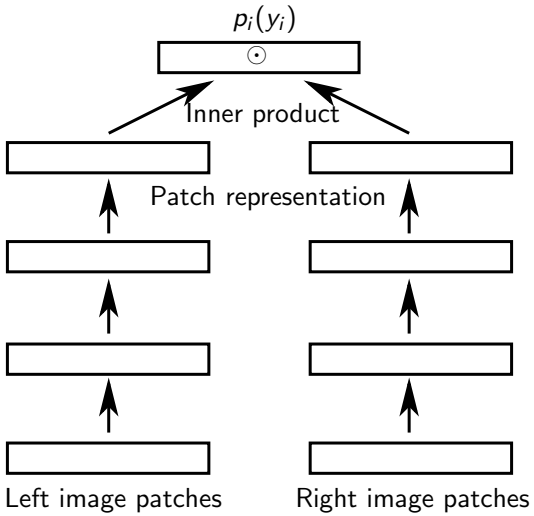


[J. Zbontar and Y. LeCun, CVPR15]

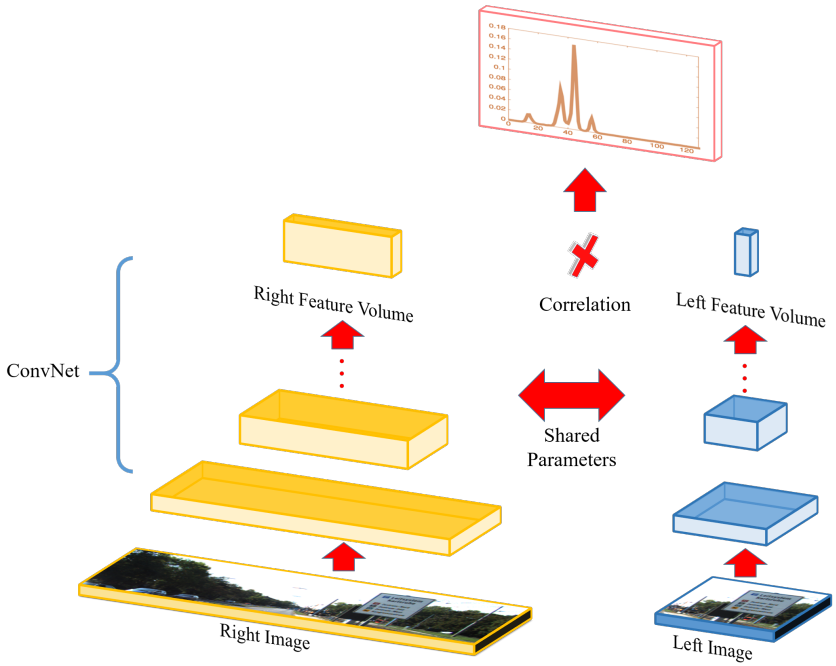
- Current approaches use a siamese network
- Combine the two branches via concatenation follow by further processing
- Treat the problem as classification (i.e., given a left and right image patches, are they a true match?)
- Too slow: 1 minute of computation on the GPU for KITTI!
- Matching not great, as scores are not correlated for different disparities

Stereo Estimation

- We propose a siamese architecture with a simple product layer, which is much faster (i.e., less than 1s in GPU)
- Train the network with multi-class loss so that the scores are calibrated, incorporating context information



(Architecture)



(Learning)

Quantitative Results on KITTI 2015

- Our approach produces much more accurate matches, 2-orders of magnitude faster than competing approaches [Zbontar & LeCun, CVPR 2015]

	$>$ 2 pixel		$>$ 3 pixel		$>$ 4 pixel		$>$ 5 pixel		End-Point		Runtime(s)
	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	
MC-CNN-acrt	15.20	16.83	12.45	14.12	11.04	12.72	10.13	11.80	4.01 px	4.66 px	22.76
Ours(37)	**9.96**	**11.67**	**7.23**	**8.97**	**5.89**	**7.62**	**5.04**	**6.78**	**1.84** px	**2.56** px	0.34

Quantitative Results on KITTI 2015

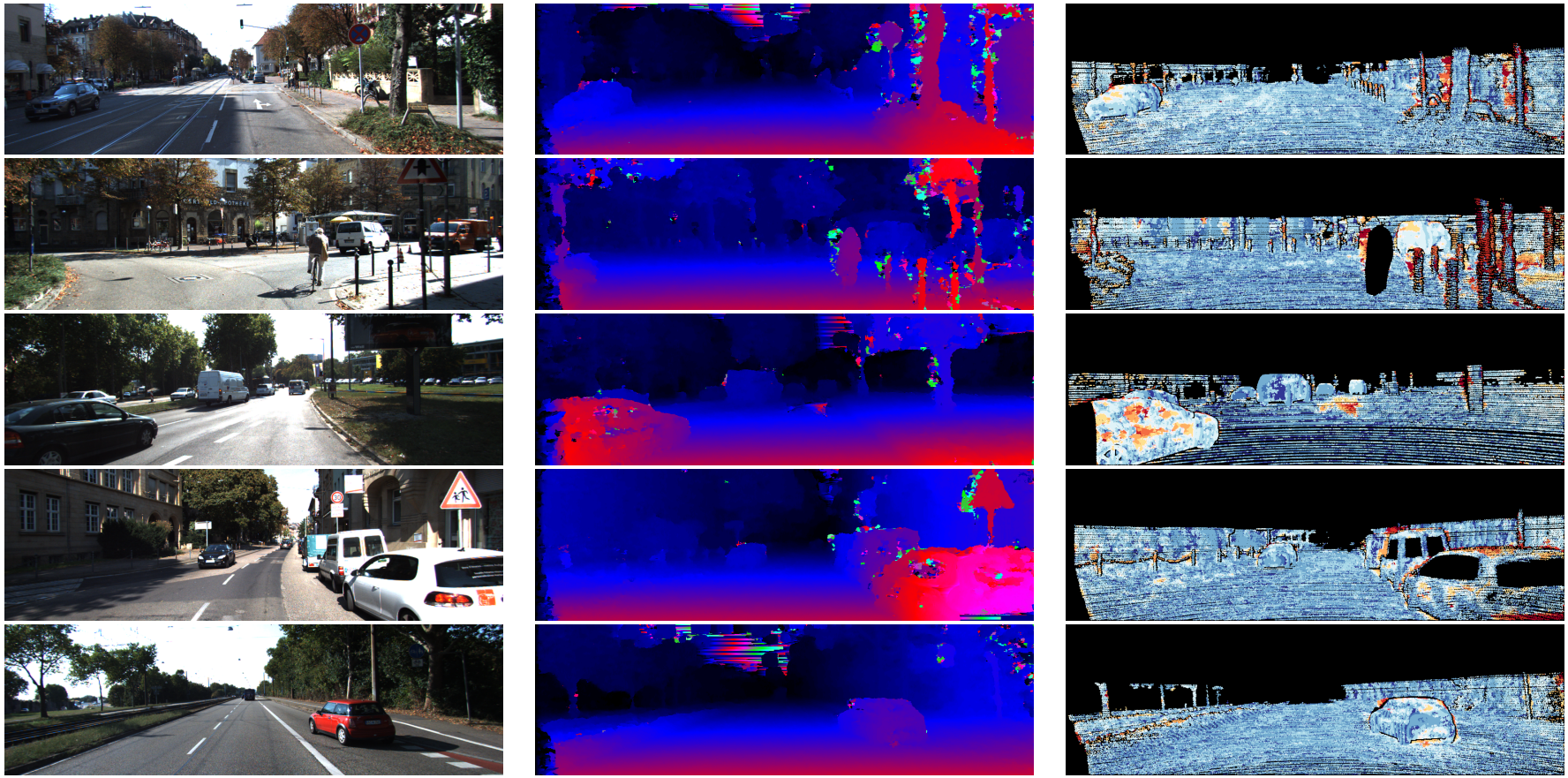
- Our approach produces much more accurate matches, 2-orders of magnitude faster than competing approaches [Zbontar & LeCun, CVPR 2015]

	> 2 pixel		> 3 pixel		> 4 pixel		> 5 pixel		End-Point		Runtime(s)
	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	Non-Occ	All	
MC-CNN-acrt	15.20	16.83	12.45	14.12	11.04	12.72	10.13	11.80	4.01 px	4.66 px	22.76
Ours(37)	**9.96**	**11.67**	**7.23**	**8.97**	**5.89**	**7.62**	**5.04**	**6.78**	**1.84** px	**2.56** px	0.34

- To be competitive this methods require cost-aggregation, semi-global matching follow by sophisticated post processing

	All/All			All/Est			Noc/All			Noc/Est			Runtime
	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	D1-bg	D1-fg	D1-all	(s)
MBM	4.69	13.05	6.08	4.69	13.05	6.08	4.33	12.12	5.61	4.33	12.12	5.61	0.13
SPS-St	3.84	12.67	5.31	3.84	12.67	5.31	3.50	11.61	4.84	3.50	11.61	4.84	2
MC-CNN	2.89	8.88	3.89	2.89	8.88	3.88	2.48	7.64	3.33	2.48	7.64	3.33	67
Displets v2	3.00	5.56	3.43	3.00	5.56	3.43	2.73	4.95	3.09	2.73	4.95	3.09	265
Ours(37)	3.73	8.58	4.54	3.73	8.58	4.54	3.32	7.44	4.00	3.32	7.44	4.00	1

Qualitative Results on KITTI 2015



- Our code is available at: http://www.cs.toronto.edu/deepLowLevelVision/