

End-to-End Training of Hybrid CNN-CRF Models for Stereo

Patrick Knöbelreiter¹ Christian Reinbacher¹ Alexander Shekhovtsov¹ Thomas Pock^{1,2}
{knoebelreiter, reinbacher, shekhovtsov, pock}@icg.tugraz.at

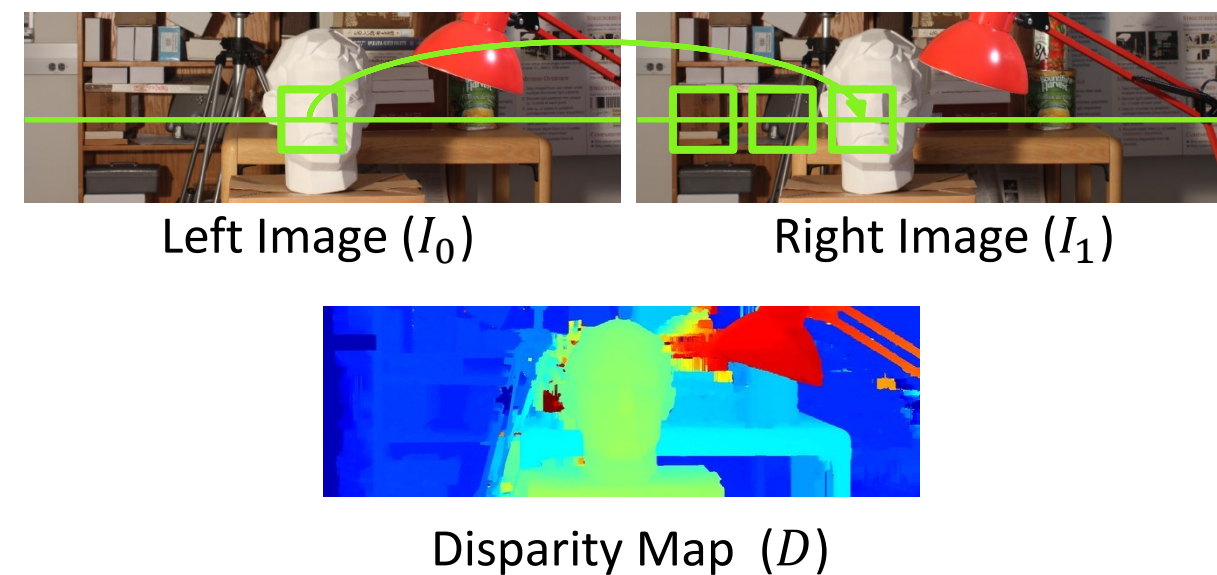
Abstract

Stereo

Find the disparity for each pixel in an image

Standard approach:

1. Densely extract features (engineered or learned) from I_0 and I_1
2. Match features for a range of disparities
3. Apply (engineered) post-processing



Motivation

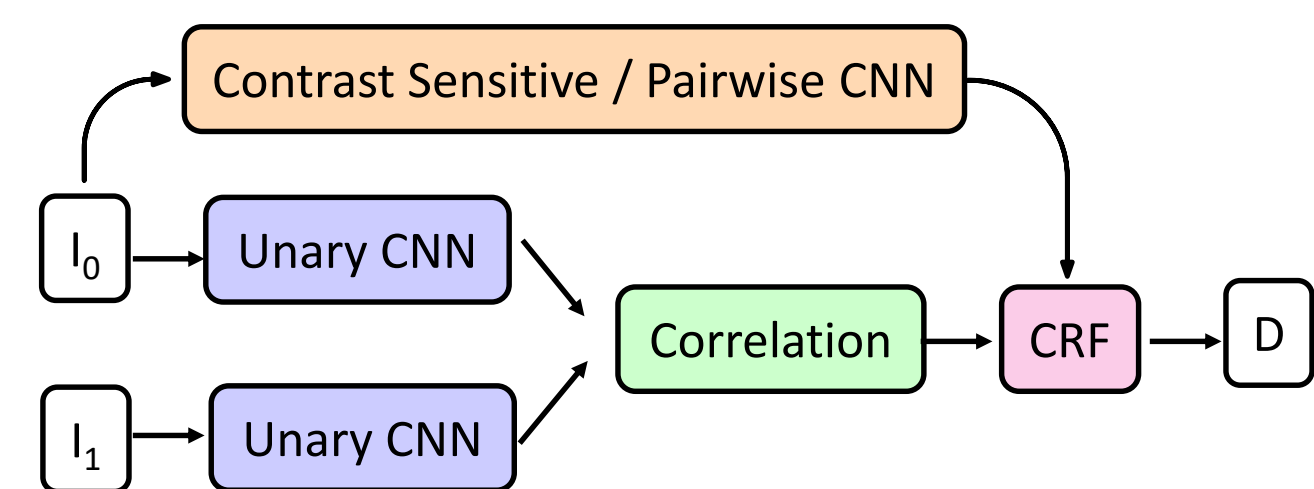
Observation:

- Recent learning-based approaches heavily rely on engineered post-processing
- ⇒ A lot of (hyper-)parameters to tune by hand
- Deep models

Idea:

- No engineered post-processing
- End-to-end learning of the complete model
- ⇒ Directly optimize for the final output
- Compact model

Model Overview



Unary CNN

- Generalize photometric/engineered features
- Automatically learn suitable features for matching from data

Correlation

- Fixed matching function
- Match learned features

Pairwise CNN

- Generalize commonly used contrast sensitive pairwise weights
- Automatically learn suitable pairwise weights for the CRF

CRF

- Global energy
- Only data- and smoothness-costs

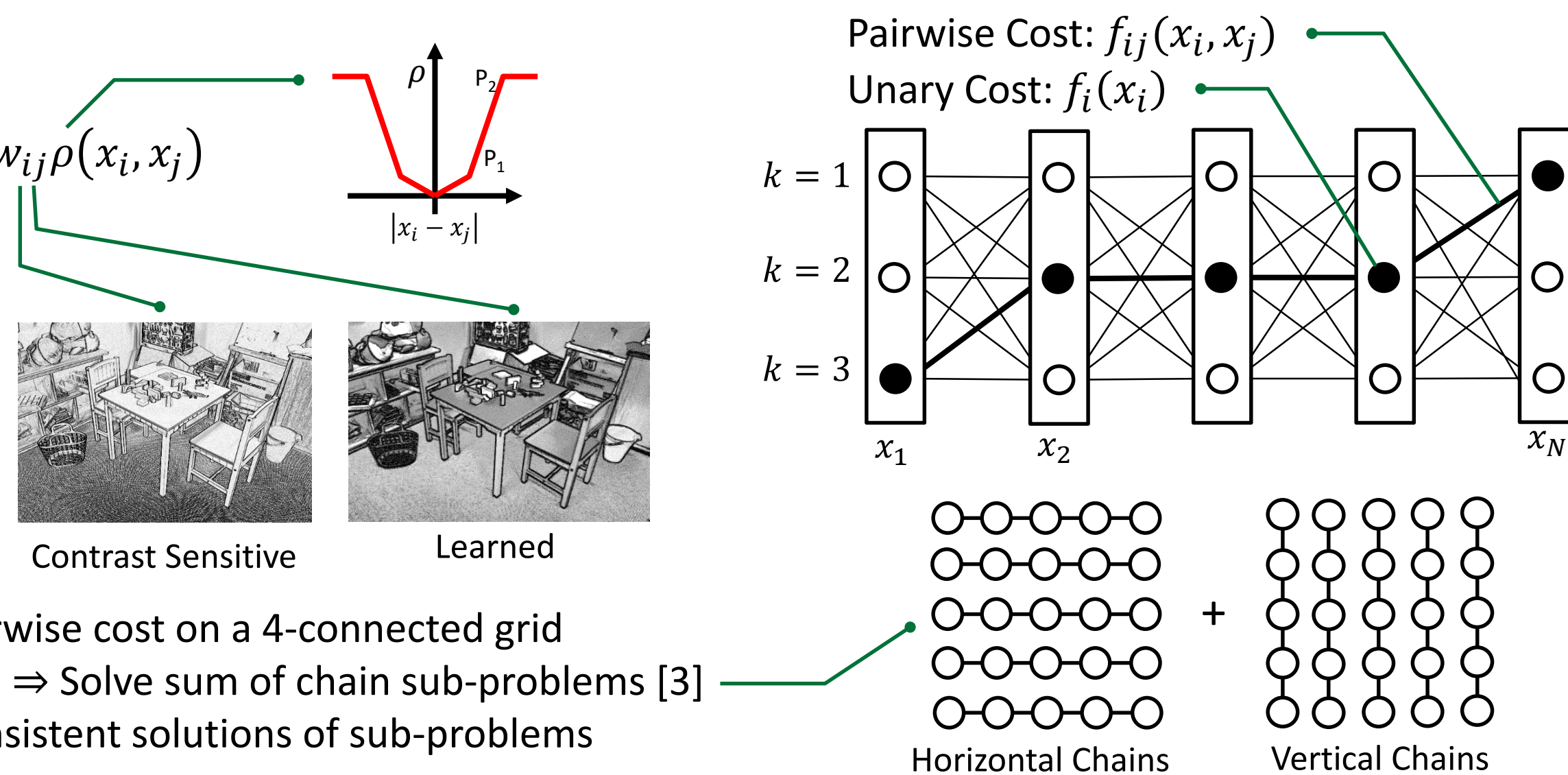
Method

Energy and Inference

$$\min_{x \in X} f(x) := \sum_{i \in V} f_i(x_i) + \sum_{ij \in E} w_{ij} \rho(x_i, x_j)$$

$$f_i(x_i) = -\sigma(\langle \phi_i^0, \phi_{i-d}^1 \rangle)$$

“Costs for candidate disparities”



- Optimize total of unary and pairwise cost on a 4-connected grid
- Dual Minorize Maximize (DMM) ⇒ Solve sum of chain sub-problems [3]
- Lagrange Multiplier ensures consistent solutions of sub-problems

End-to-End Training

Bi-level Optimization Problem

$$\min_{\theta} l(x, x^*) \quad \text{s.t. } x \in \arg \min_{x \in X} f(x)$$

“Learn parameters θ of CNNs, such that the minimizer of the CRF model minimizes a certain loss function”

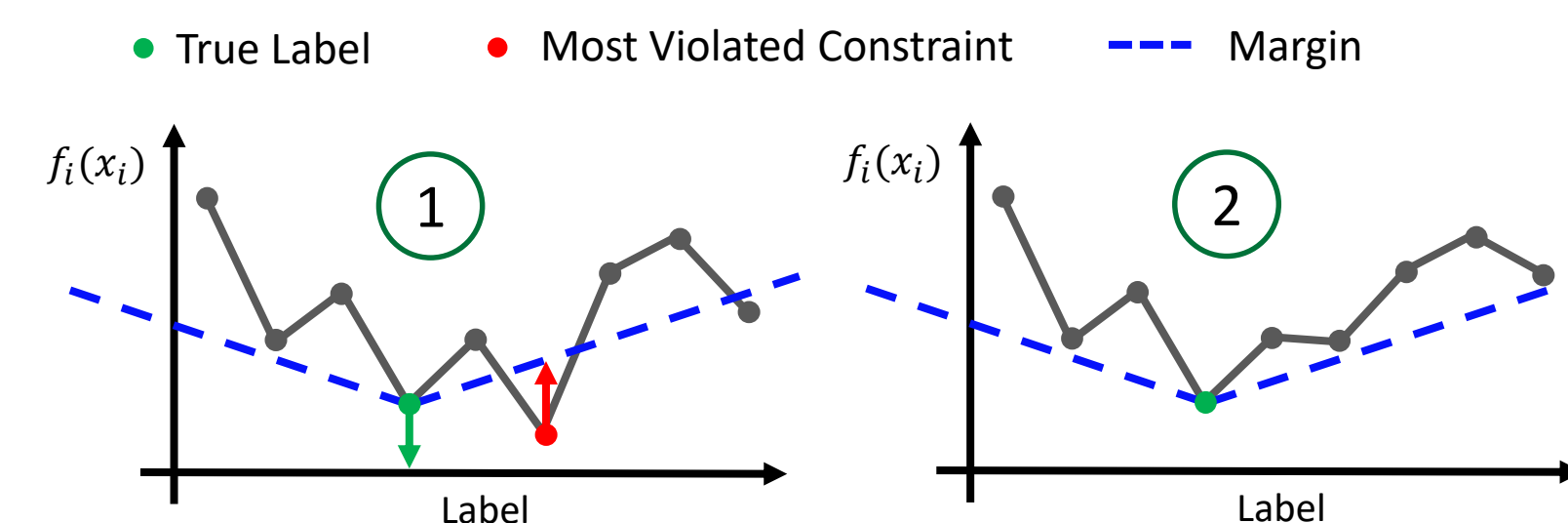
- Directly computing a gradient is not possible, because $\arg \min$ is not differentiable
- Compute a gradient based on the structured output support vector machine (SSVM)

⇒ Minimize upper bound of the loss

Training Procedure

- (1) Find most violated constraint
- (2a) Decrease energy of true label
- (2b) Increase energy of most violated constraint

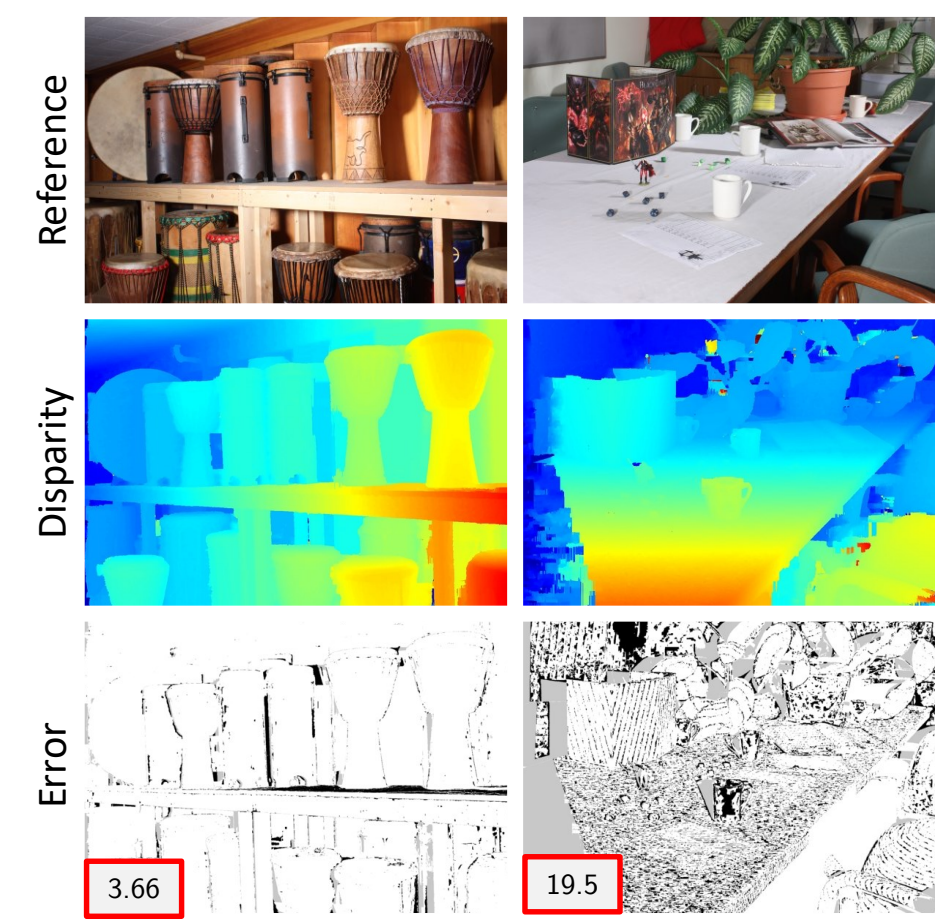
Interpretation



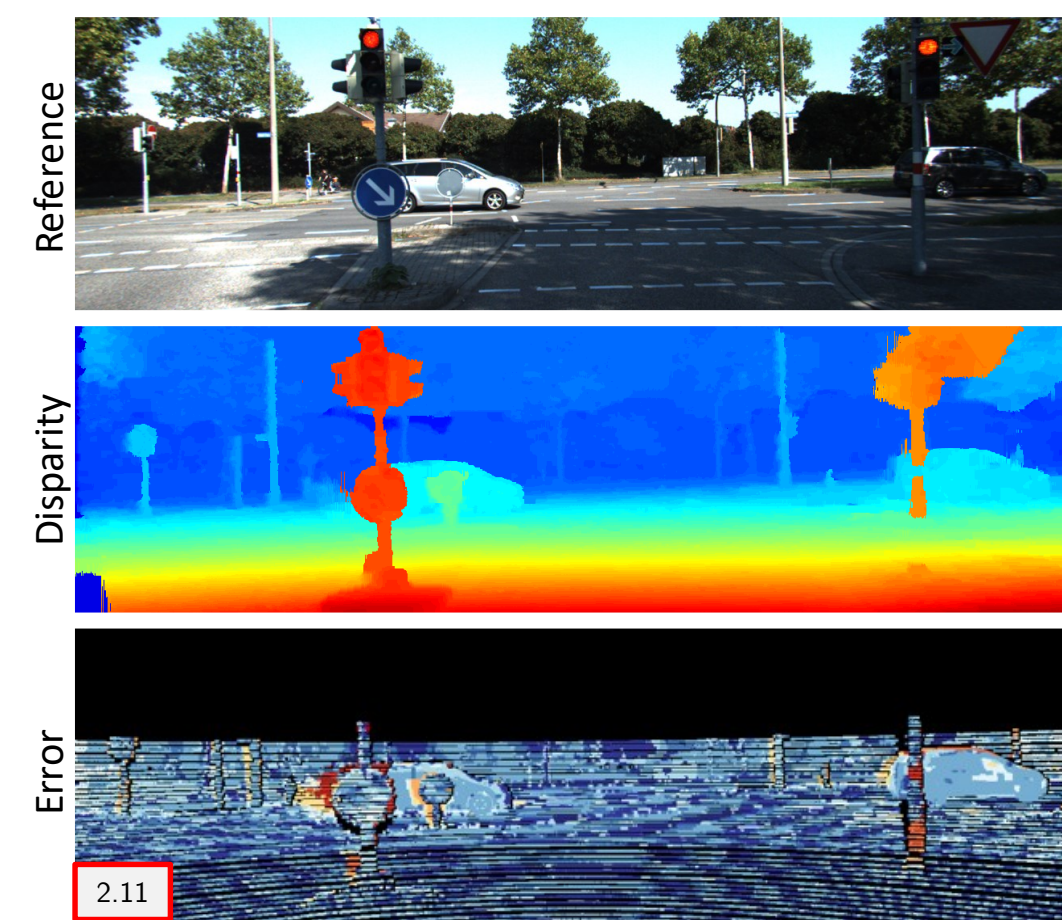
Results

Qualitative Results

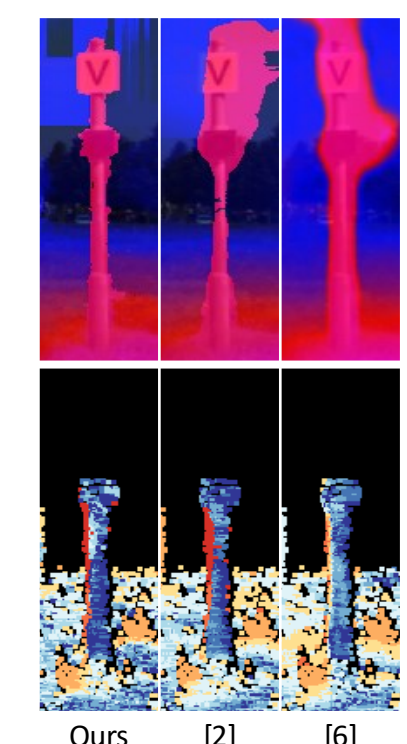
Middlebury 2014



Kitti 2015



Kitti Closeup



Benchmark Results

	Method	$\emptyset$ RMS	$\emptyset$ bad2	Time/MP	Parameters	Post-Processing
Middlebury 2014	[1]	21.3	8.29	112s	830k	CA, SGM, SE, MF, BF
	[5]	15.0	8.62	140s	830k	CA, SGM, SE, MF, BF, RBS
	Ours	14.4	12.5	3.69s	281k	-
Kitti 2015	[1]	3.33	3.89	67s	830k	CA, SGM, SE, MF, BF
	[2]	4.00	4.54	1s	700k	CA, SGM, LR, SE, MF, BF, RBS
	[6]	4.32	4.34	0.06s	42M	-
	Ours	4.84	5.50	1.3s	281k	-

CA...Cost Aggregation, SGM...Semi-Global Matching, SE...Sublabel Enhancement, LR...Left-Right Check, MF...Median Filtering, BF...Bilateral Filtering, RBS...Robust Bilateral Solver

References

- [1] Zbontar, J. et al. Stereo matching by training a convolutional neural network to compare patches. *JMLR* 2016.
- [2] Luo, W. et al. Efficient deep learning for stereo matching. *CVPR* 2016.
- [3] Shekhovtsov, A. et al. Solving dense image matching in real-time using discrete-continuous optimization. In *CVWW* 2016.
- [4] Taskar, B. et al. Max-margin markov networks. *MIT Press* 2003.
- [5] Barron, J. T. et al. The fast bilateral solver. *ECCV* 2016
- [6] Mayer, N. et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. *CVPR* 2016

Conclusion

- Principled approach without post-processing - only optimization
- Accurate and fast
- Exploit GPU based CRF solver
- Learn unary and pairwise term such that the CRF works best
- All parameters learned from data
- Code online:

<https://github.com/VLOGroup/cnn-crf-stereo>

