Graz University of Technology

Best Paper Award @GCPR 2019

Learned Collaborative Stereo Refinement

Patrick Knöbelreiter and Thomas Pock

Institute for Computer Graphics and Vision
Graz University of Technology

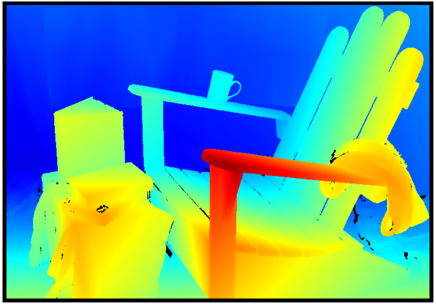
October 8^{th}, 2019

Introduction

Stereo

- **Input:** Two images from a calibrated camera pair
- **Task:** For each pixel in the left image find the corresponding pixel in the right image

Left/Right Input [Middlebury 2014]



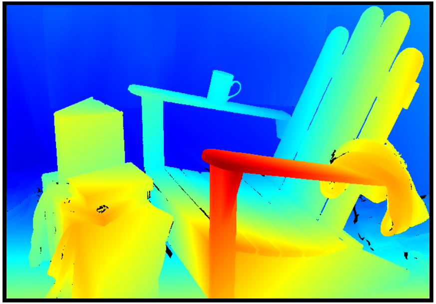
Disparity Map

Introduction

Stereo

- **Input:** Two images from a calibrated camera pair
- **Task:** For each pixel in the left image find the corresponding pixel in the right image

Left/Right Input [Middlebury 2014]



Disparity Map

Introduction

Stereo Pipeline [Scharstein and Szeliski]

Matching Cost Computation → Cost Aggregation → Disparity Computation → Disparity Refinement

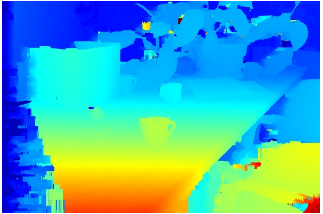
Reference Image [Middlebury]

Introduction

Stereo Pipeline [Scharstein and Szeliski]

Matching Cost Computation → Cost Aggregation → Disparity Computation → Disparity Refinement

Reference Image [Middlebury]



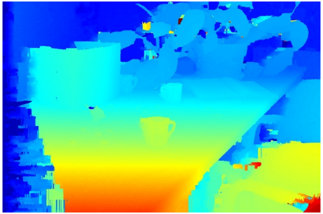
CNN-CRF [CVPR'17]

Introduction

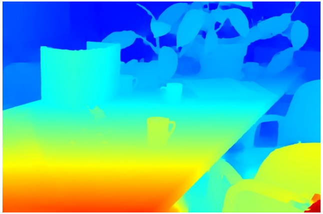
Stereo Pipeline [Scharstein and Szeliski]

Matching Cost Computation → Cost Aggregation → Disparity Computation → **Disparity Refinement**

Reference Image [Middlebury]



CNN-CRF [CVPR'17]

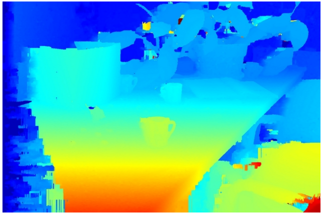


Refined Result (ours)

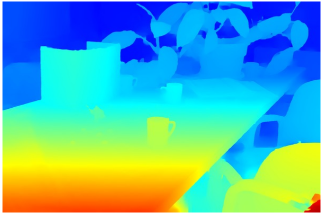
Introduction

Disparity Refinement

Reference Image



CNN-CRF Result



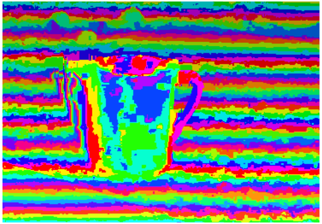
Refined Result (ours)

Introduction

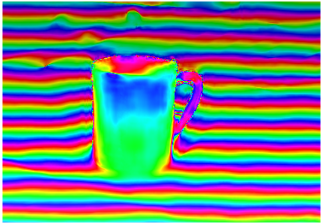
Disparity Refinement

- Reduce discretization artifacts
- Handle slanted surfaces correctly
- Get sub-pixel accuracy

Reference Image



CNN-CRF Result



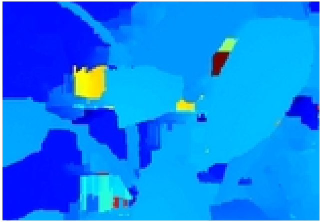
Refined Result (ours)

Introduction

Disparity Refinement

- Reduce discretization artifacts
- Handle slanted surfaces correctly
- Get sub-pixel accuracy
- Reduce artifacts in occlusions

Reference Image



CNN-CRF Result

Refined Result (ours)

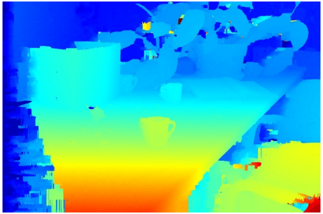
Introduction

Disparity Refinement

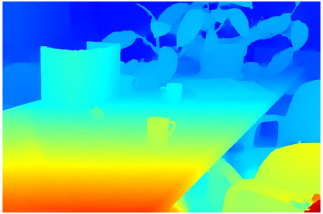
- Reduce discretization artifacts
- Handle slanted surfaces correctly
- Get sub-pixel accuracy
- Reduce artifacts in occlusions

How can we do this?

Reference Image



CNN-CRF Result



Refined Result (ours)

Related Work - Stereo Refinement

Optimization based

- Variational Methods
- Total Variation (TV) [Ranftl *et al.*, Kuschk *et al.*]
- Higher order variant of TV [Ranftl *et al.*]
- Learned Reaction Diffusion [Chen *et al.*]
- The Fast Bilateral Solver [Barron and Poole]

Related Work - Stereo Refinement

Optimization based

- Variational Methods
- Total Variation (TV) [Ranftl *et al.*, Kuschk *et al.*]
- Higher order variant of TV [Ranftl *et al.*]
- Learned Reaction Diffusion [Chen *et al.*]
- The Fast Bilateral Solver [Barron and Poole]

Learning based

- Cascade Residual Learning [Pang *et al.*]
  Apply the same network twice
- StereoNet [Khamis *et al.*]
  Disparity refinement with residual blocks
- Learning for disparity estimation through feature consistency [Liang *et al.*]
- Detect, Replace, Refine [Gidaris and Komodakis]

Related Work - Stereo Refinement

Optimization based

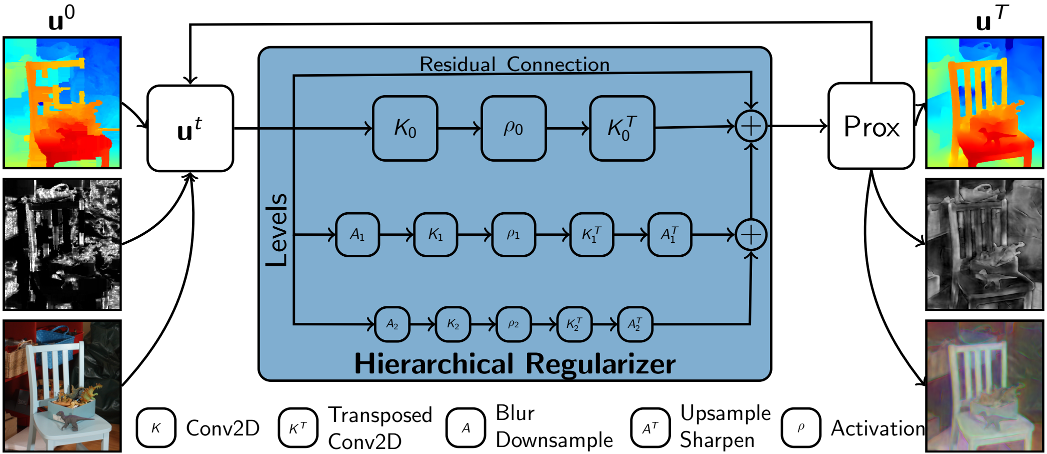
- Variational Methods
- Total Variation (TV) [Ranftl *et al.*, Kuschk *et al.*]
- Higher order variant of TV [Ranftl *et al.*]
- Learned Reaction Diffusion [Chen *et al.*]
- The Fast Bilateral Solver [Barron and Poole]

Learning based

- Cascade Residual Learning [Pang *et al.*]
  Apply the same network twice
- StereoNet [Khamis *et al.*]
  Disparity refinement with residual blocks
- Learning for disparity estimation through feature consistency [Liang *et al.*]
- Detect, Replace, Refine [Gidaris and Komodakis]

Combine the structure of optimization and the power of deep learning

Model Overview



Collaborative Stereo Refinement

Optimization Problem

$$\min_{\mathbf{u}} \underbrace{\mathcal{R}(\mathbf{u})}_{Regularizer} + \underbrace{\mathcal{D}(\mathbf{u})}_{Data\,Term}, \quad \mathbf{u} = (\mathbf{u}^{rgb}, u^d, u^c) : \Omega \to \mathbb{R}^5, \quad \Omega \subset \mathbb{N}_+^2$$

Regularizer

- Variant of Field of Experts (FoE) [Roth *et al.*]

$$\mathcal{R}(\mathbf{u};\theta) = \sum_{l=1}^{L} \sum_{k=1}^{K} \sum_{x \in \Omega} \phi_k^l \left(\left(K_k^l A^l \mathbf{u} \right)(x) \right)$$

Data Term

$$\mathcal{D}(\mathbf{u};\theta) = \frac{\lambda}{2} \|\mathbf{u}^{rgb} - \mathbf{f}^0\|^2 + \mu \|u^c - c\|_1 + \nu \|u^d - \check{d}\|_{u^c,1}$$

Collaborative Stereo Refinement

Variational Network

- Unrolling the iterates of the Proximal Gradient Method yields the Variational Network:

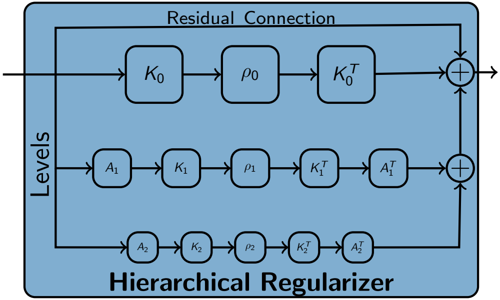
$$\mathbf{u}_{t+1} = \mathrm{prox}_{\alpha_t \mathcal{D}(\cdot,\theta_t)}(\mathbf{u}_t - \alpha_t \nabla \mathcal{R}(\mathbf{u}_t, \theta_t)),$$

with

$$0 \leq t \leq T-1$$

and

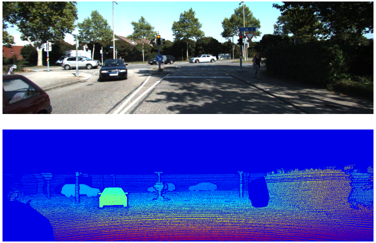
$$\nabla \mathcal{R}(\mathbf{u}) = \sum_{l=1}^{L} \sum_{k=1}^{K} (K_k^l A^l)^T \rho_k^l \left(K_k^l A^l \mathbf{u}\right)$$



Learning

Datasets

- Middlebury 2014:
  Indoor, high-resolution, dense ground-truth
- Kitti 2015:
  Outdoor, sparse ground-truth



[Kitti 2015]

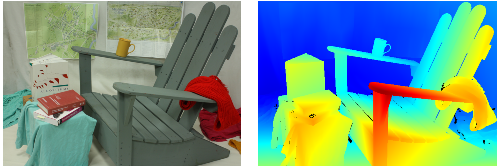
Training

- Truncated Huber loss

$$\min_{\theta \in \Theta} \sum_{s=1}^{S} \sum_{i=1}^{MN} \min \left(|u^{d}_{s,T}(x, \theta) - d^{*}_{s}(x)|_{\delta},\ \tau \right)$$

$$|r|_{\delta} = \begin{cases} \frac{r^2}{2\delta} & \text{if } |r| \leq \delta \\ |r| - \frac{\delta}{2} & \text{else} \end{cases}$$

- Adam optimizer with learning rate 10^{-3}



[Middlebury 2014]

Experiments

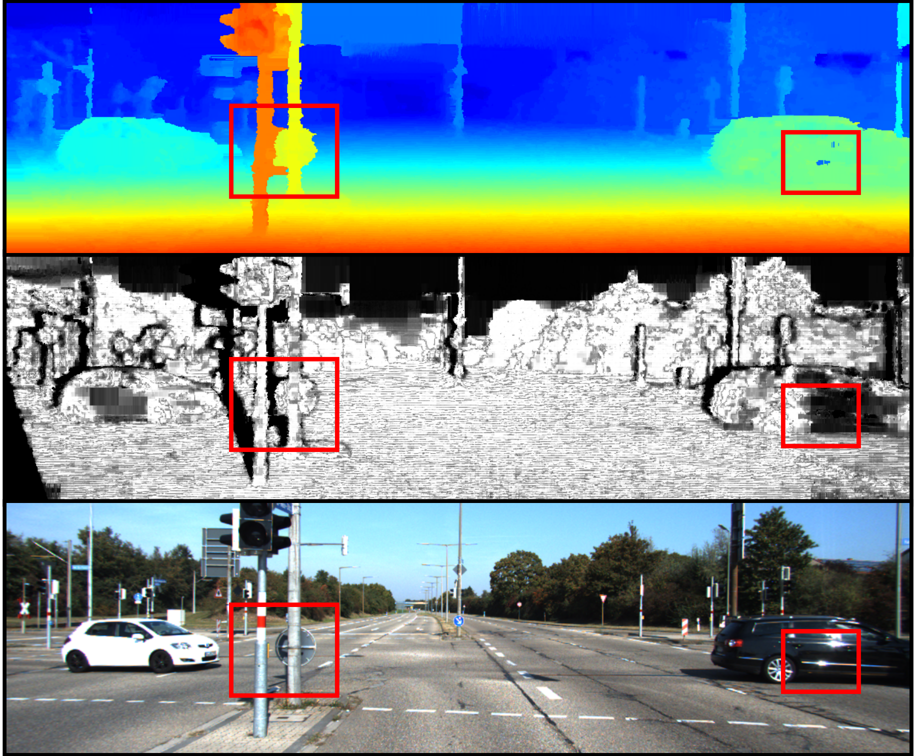
Ablation Study

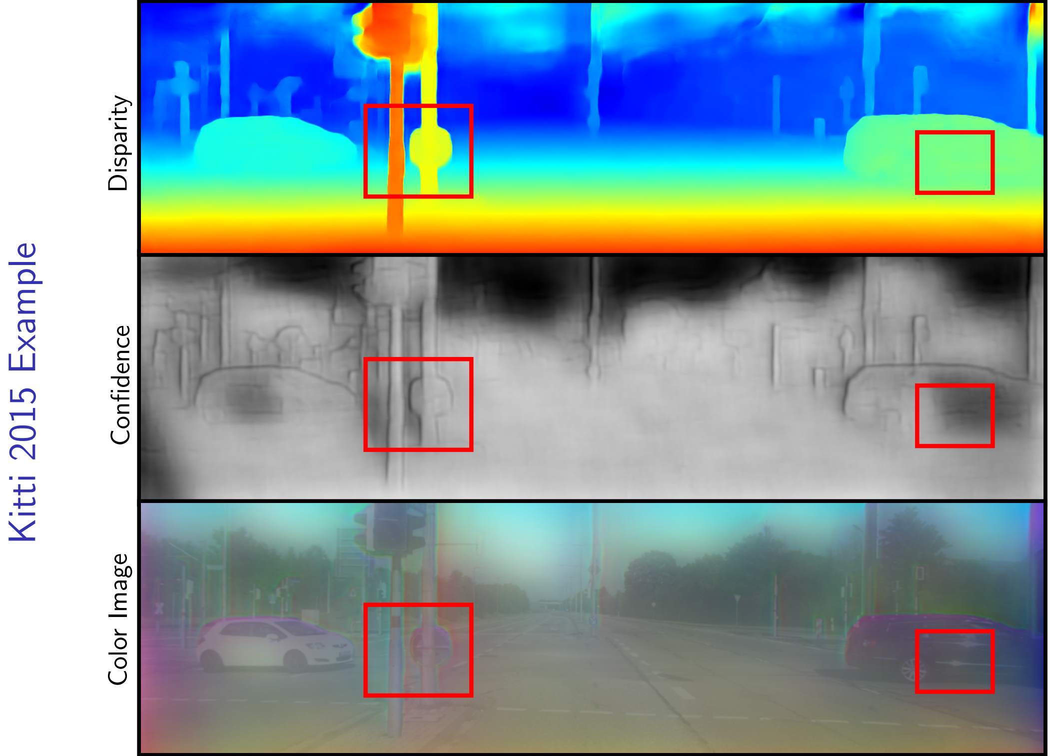
- Refine the winner-takes-all (WTA) result of a UNet [Long *et al.*]
- $VN_4^{7,5}$ is a variational refinement network with 7 steps, 4 levels and 5×5 filters
- bad3 $=$ percentage of pixels with an error $\leq$ 3px

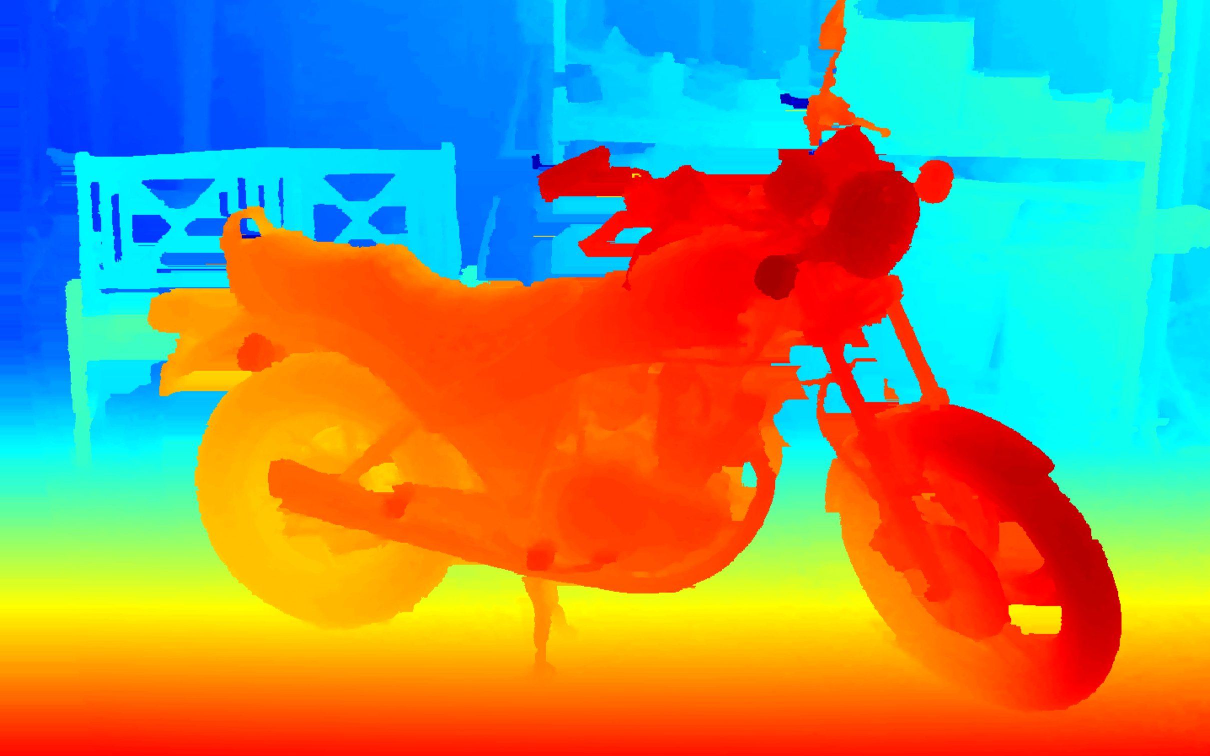
Model	WTA	FBF	$VN_4^{7,5}$	$VN_4^{7,5}$	$VN_4^{7,5}$	$VN_4^{7,5}$	$VN_4^{7,5}$	$VN_4^{7,5}$	$VN_3^{5,7}$	$VN_2^{8,7}$	$VN_4^{14,3}$	$VN_5^{11,3}$
Conf		✓		✓	✓		✓	✓	✓	✓	✓	✓
Img		✓		✓		✓	✓	✓	✓	✓	✓	✓
Occlp		✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Joint								✓	✓	✓	✓	✓
[bad3] occ	8.24	7.48	5.42	5.12	4.43	3.77	3.46	**3.37**	3.43	3.62	4.37	4.25
noc	6.78	6.08	4.68	3.98	3.90	3.07	2.72	**2.55**	2.58	2.97	3.71	3.49

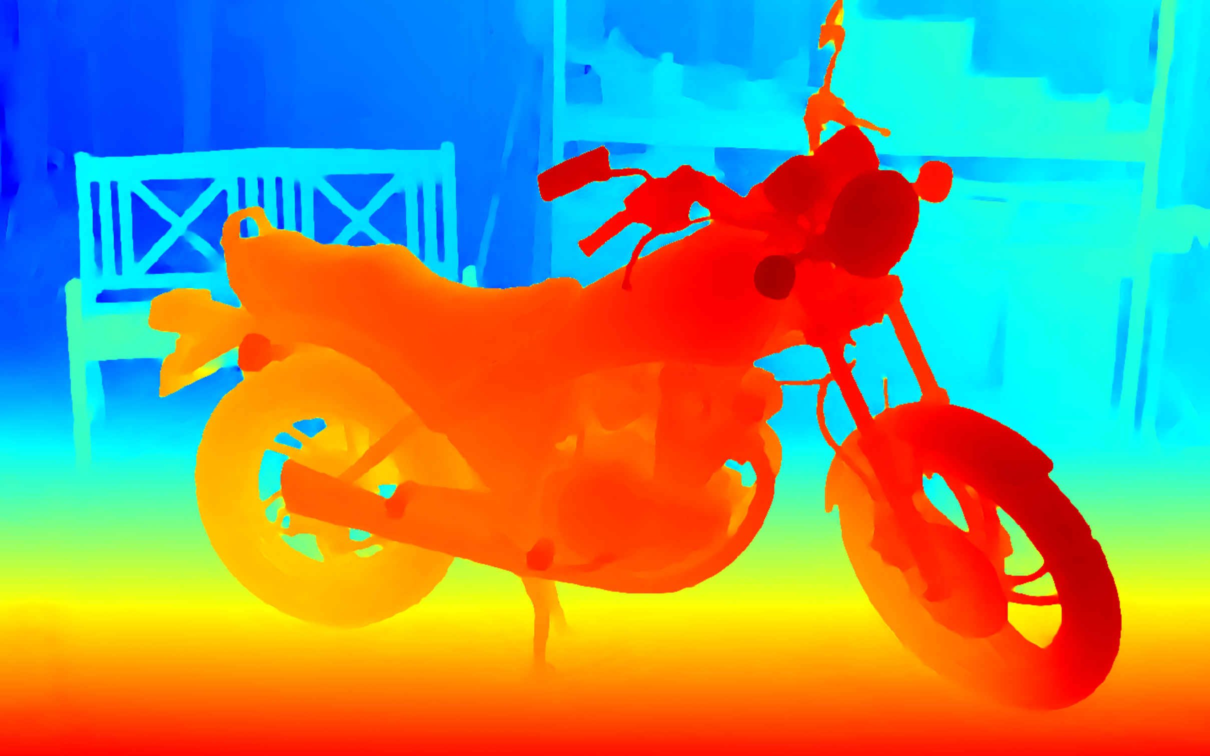
Table: Ablation Study on the Kitti 2015 dataset

Kitti 2015 Example




Kitti 2015 Example
Disparity
Confidence
Color Image

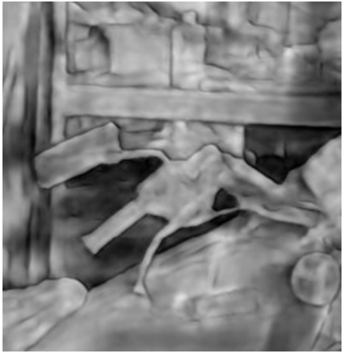




Experiments - Qualitative Results

Middlebury

Disparity



Confidence

Color

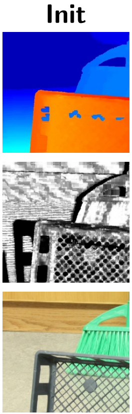
Experiments

Public Benchmarks

	Method	Kitti 2015 noc	Kitti 2015 all	Kitti 2015 ∅R	Middlebury 2014 bad0.5	bad1	bad2	bad4	avg	rms	time	∅R
Training Set	PSMNet [Chang *et al.*]	-	**1.83**	-	90.0	78.1	58.5	32.2	9.60	21.7	2.62	44
	PDS [Tulyakov *et al.*]	-	-	-	54.2	26.1	11.4	5.10	1.98	9.10	10	8
	MC-CNN [Zbontar and LeCun]	-	-	-	42.1	20.5	11.7	7.94	3.87	16.5	1.26	9
	CNN-CRF [Knöbelreiter *et al.*]	-	4.04	-	56.1	25.1	10.8	6.12	2.30	9.89	3.53	10
	CNN-CRF + VN (ours)	**1.90**	*2.04*	-	**41.8**	**17.1**	**7.05**	**2.96**	**1.21**	**5.80**	4.06	*2*
Test Set	PSMNet [Chang *et al.*]	**2.14**	**2.32**	17	81.1	63.9	42.1	23.5	6.68	19.4	2.62	33
	PDS [Tulyakov *et al.*]	2.36	2.58	19	58.9	21.1	14.2	6.98	3.27	15.7	10.3	9
	MC-CNN [Zbontar and LeCun]	3.33	3.89	32	**41.3**	**18.0**	**9.47**	6.7	4.37	22.4	1.26	6
	CNN-CRF [Knöbelreiter *et al.*]	4.84	5.50	36	60.9	31.9	12.5	**6.61**	3.02	14.4	3.53	8
	CNN-CRF + VN (ours)	*4.45*	*4.85*	*33*	*56.2*	*30.0*	14.2	7.71	**2.49**	**10.8**	4.06	*6*

Experiments

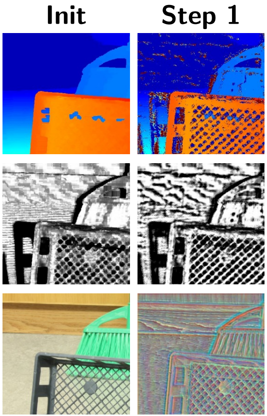
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

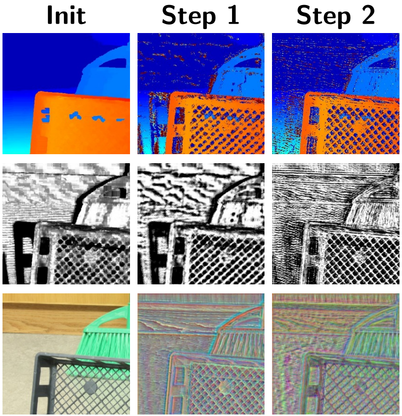
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

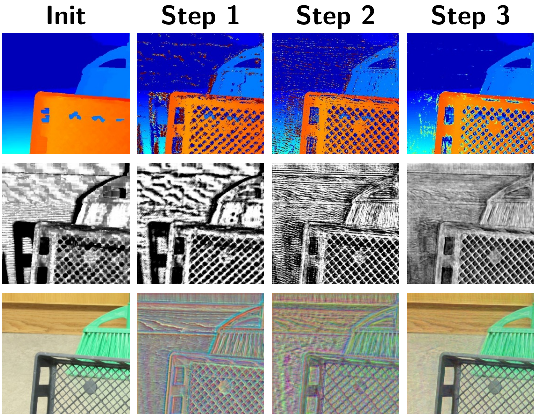
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

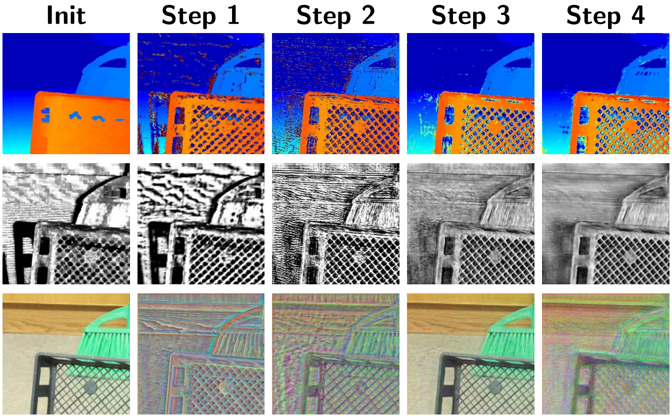
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

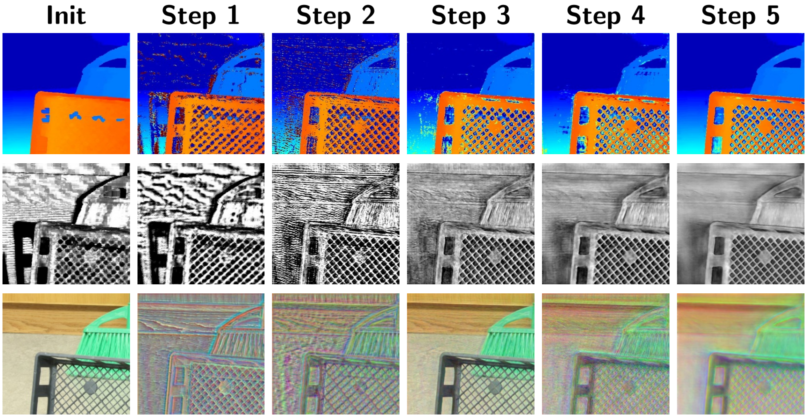
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

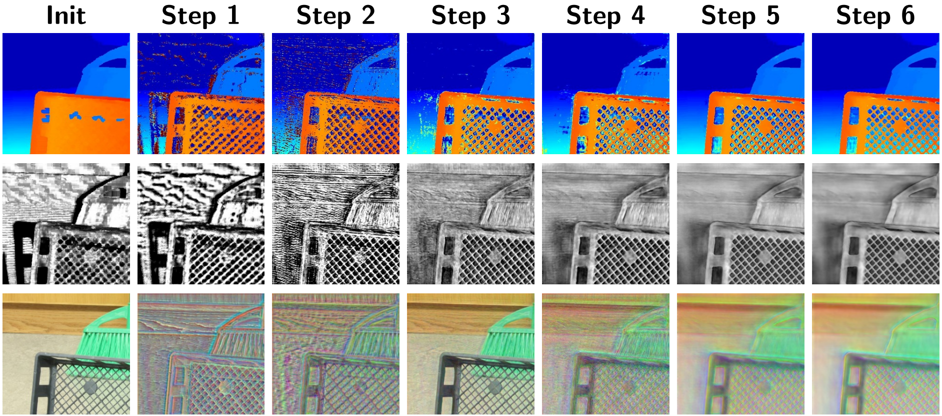
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

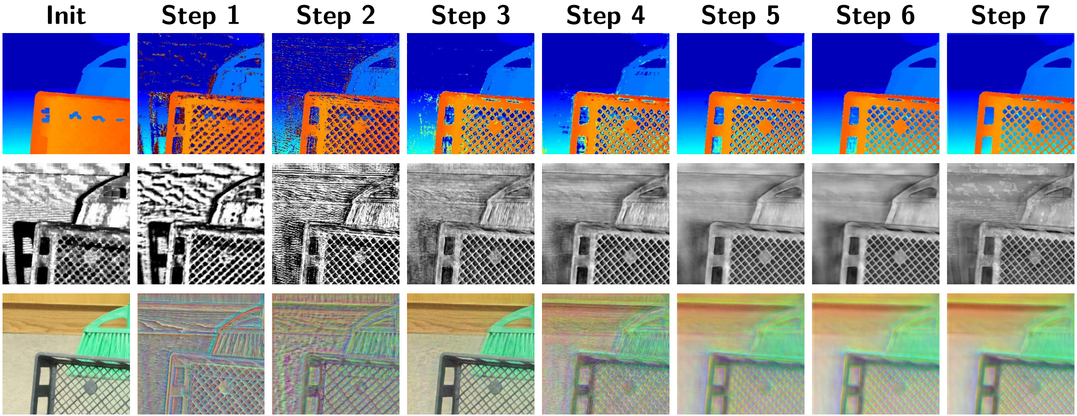
Visualization of the Refinement Process



Example from the Middlebury dataset

Experiments

Visualization of the Refinement Process

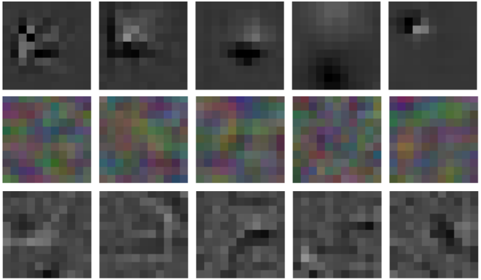


Example from the Middlebury dataset

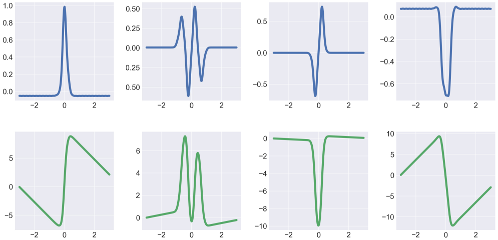
Experiments

Visualizing the Model

- Learned filters contain structure
- Activation functions converge to smooth functions
- Both are thus easily interpretable



(a) Filters



(b) Activation/Potential Function

Conclusion

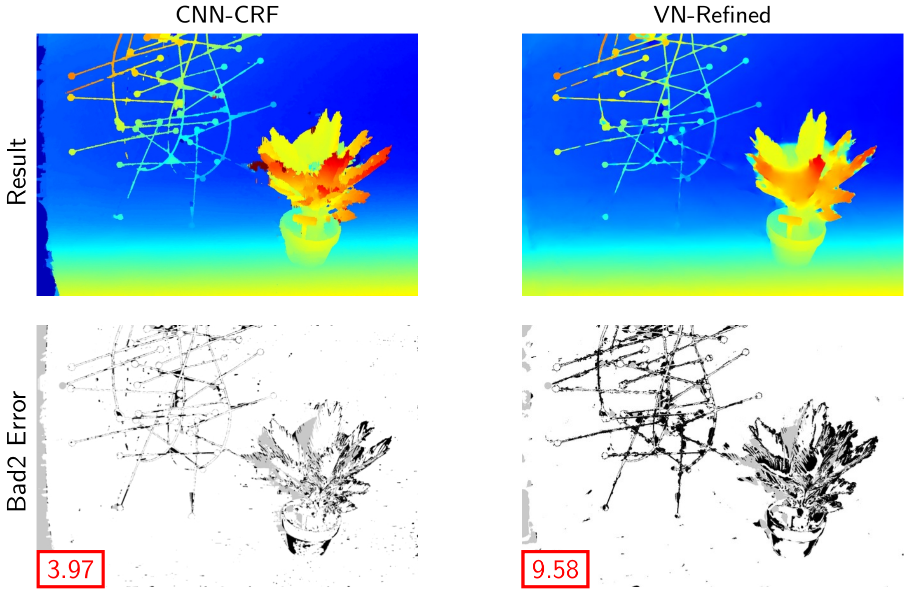
- Stereo refinement module
  - exploiting color, confidence and disparity map
  - operating on multiple spatial resolutions
  - that arises naturally from the iterates of a PGM
- Interpretable steps, filters and activation functions
- Can be used on top of existing stereo methods
- Good and bad is often a question of the used metric

Conclusion

- Stereo refinement module
  - exploiting color, confidence and disparity map
  - operating on multiple spatial resolutions
  - that arises naturally from the iterates of a PGM
- Interpretable steps, filters and activation functions
- Can be used on top of existing stereo methods
- Good and bad is often a question of the used metric

Thank you for your attention!

Experiments - Benchmark Results



Experiments - Benchmark Results