Middlebury

PerMeCoP2015

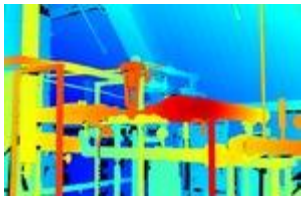
The Middlebury Stereo Evaluation Version 3

Daniel Scharstein

Department of Computer Science
Middlebury College





This work is supported by NSF grants IIS-0413169 and IIS-0917109

Joint work with …

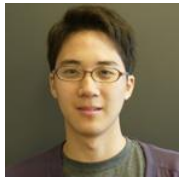
Heiko Hirschmüller, DLR / roboception

Duncan Levear ’15

Nera Nesic ’13

Greg Krathwohl ’14

Xi Wang ’14



York Kitajima ’15

Porter Westling ’12

…building on joint work with

Rick Szeliski, Microsoft Research

Alan Lim ’09

Anna Blasiak ’07

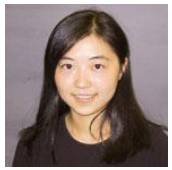
Sarri Al-Nashashibi ’08

Jeff Wehrwein ’08

Brad Hiebert-Treuer ’07

Gonzalo Alonso ’06

Alexander Vandenberg-Rodes



Jiaxin (Lily) Fu ’03

Padma Ugbabe ’03

Motivation

- StereoEval v.2 unchanged since 2005, virtually solved, overfitting
- KITTI (2012) was important addition, but only provides partial new challenges
- Increasing need for more datasets

Goals

- Provide new hi-res datasets
  to propel research in stereo matching

- Address common problems with existing benchmarks

- Provide useful UI

Middlebury Stereo Page

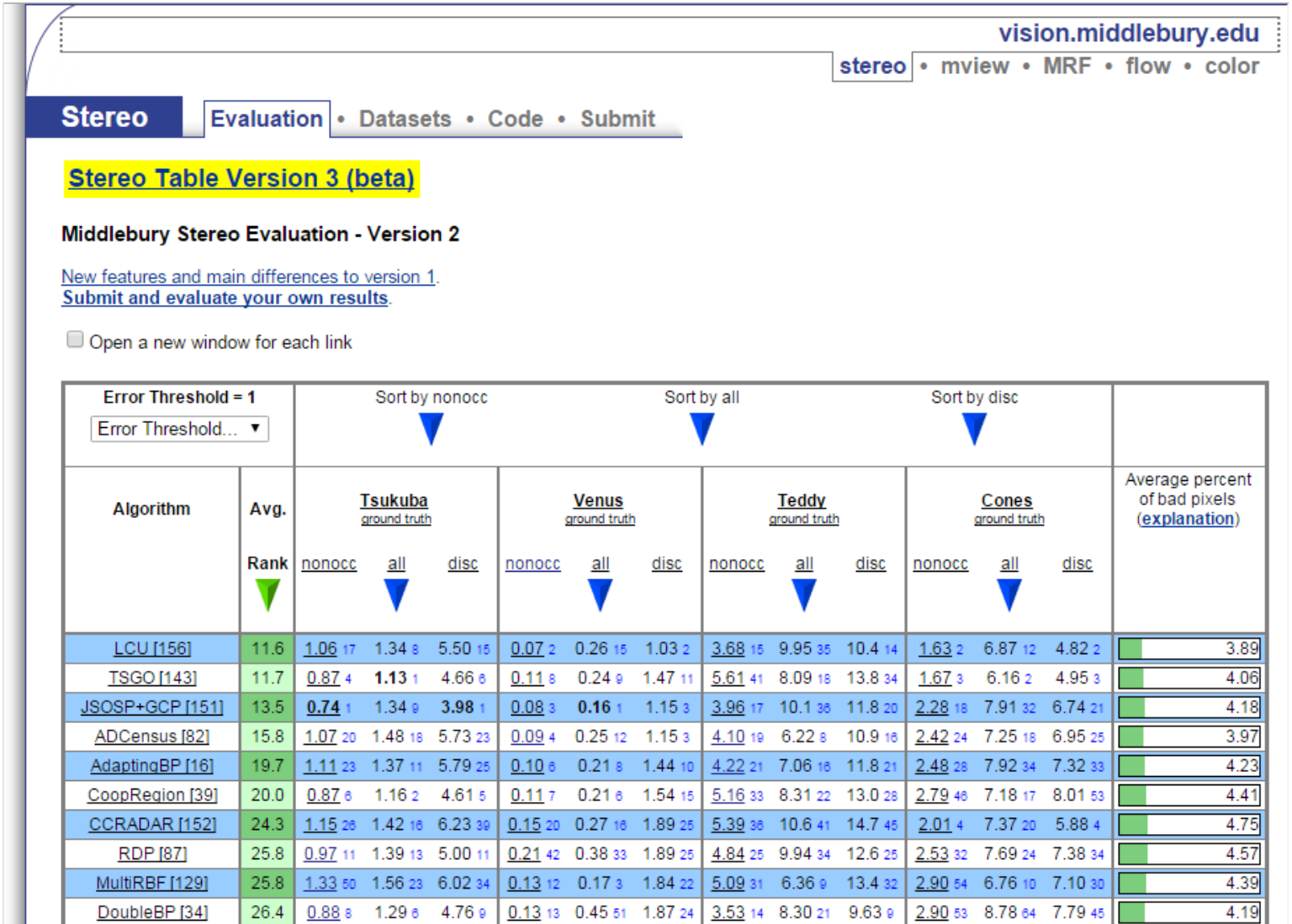
(Scharstein & Szeliski – CVPR 2001, IJCV 2002)

- vision.middlebury.edu/stereo
- Evaluator with web interface

v.1 (2002) by Lily Fu ’03 v.2 (2005) by Anna Blasiak ’07

Middlebury stereo benchmark v.2

Currently 162 entries



vision.middlebury.edu

stereo • mview • MRF • flow • color

Stereo | Evaluation • Datasets • Code • Submit

Stereo Table Version 3 (beta)

Middlebury Stereo Evaluation - Version 2

New features and main differences to version 1.
Submit and evaluate your own results.

Open a new window for each link

Error Threshold = 1 (Error Threshold...) | Sort by nonocc | Sort by all | Sort by disc

Algorithm	Avg. Rank	Tsukuba ground truth			Venus ground truth			Teddy ground truth			Cones ground truth			Average percent of bad pixels (explanation)
		nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	nonocc	all	disc	
LCU [156]	11.6	1.06 17	1.34 8	5.50 15	0.07 2	0.26 15	1.03 2	3.68 15	9.95 35	10.4 14	1.63 2	6.87 12	4.82 2	3.89
TSGO [143]	11.7	0.87 4	**1.13** 1	4.66 6	0.11 8	0.24 9	1.47 11	5.61 41	8.09 18	13.8 34	1.67 3	6.16 2	4.95 3	4.06
JSOSP+GCP [151]	13.5	**0.74** 1	1.34 9	**3.98** 1	0.08 3	**0.16** 1	1.15 3	3.96 17	10.1 36	11.8 20	2.28 18	7.91 32	6.74 21	4.18
ADCensus [82]	15.8	1.07 20	1.48 18	5.73 23	0.09 4	0.25 12	1.15 3	4.10 19	6.22 8	10.9 16	2.42 24	7.25 18	6.95 25	3.97
AdaptingBP [16]	19.7	1.11 23	1.37 11	5.79 25	0.10 6	0.21 8	1.44 10	4.22 21	7.06 16	11.8 21	2.48 28	7.92 34	7.32 33	4.23
CoopRegion [39]	20.0	0.87 6	1.16 2	4.61 5	0.11 7	0.21 6	1.54 15	5.16 33	8.31 22	13.0 28	2.79 46	7.18 17	8.01 53	4.41
CCRADAR [152]	24.3	1.15 26	1.42 16	6.23 39	0.15 20	0.27 16	1.89 25	5.39 38	10.6 41	14.7 45	2.01 4	7.37 20	5.88 4	4.75
RDP [87]	25.8	0.97 11	1.39 13	5.00 11	0.21 42	0.38 33	1.89 25	4.84 25	9.94 34	12.6 25	2.53 32	7.69 24	7.38 34	4.57
MultiRBF [129]	25.8	1.33 50	1.56 23	6.02 34	0.13 12	0.17 3	1.84 22	5.09 31	6.36 9	13.4 32	2.90 54	6.76 10	7.10 30	4.39
DoubleBP [34]	26.4	0.88 8	1.29 6	4.76 9	0.13 13	0.45 51	1.87 24	3.53 14	8.30 21	9.63 9	2.90 53	8.78 64	7.79 45	4.19

Limitations of existing stereo benchmarks

	Midd v.2	KITTI stereo
# of image pairs	4	194 + 195
Image size	< 0.2 MP	< 0.5 MP
Disparity range	16 … 60	70 … 150
Scene complexity	Limited	Limited
Scene variety	Limited	Limited
Realism	Limited	Good
Radiometric challenges	None	Some
Ground truth accuracy	Good	Limited
Ground truth coverage	Good	Limited
Control against overfitting	None	Some

Comparison with Midd v.3

	Midd v.2	KITTI stereo	Midd v.3
# of image pairs	4	194 + 195	15 + 15
Image size	< 0.2 MP	< 0.5 MP	5-6 MP
Disparity range	16 … 60	70 … 150	250 … 800
Scene complexity	Limited	Limited	Good
Scene variety	Limited	Limited	Good*
Realism	Limited	Good	Good*
Radiometric challenges	None	Some	Some
Ground truth accuracy	Good	Limited	Very good
Ground truth coverage	Good	Limited	Good
Control against overfitting	None	Some	Yes

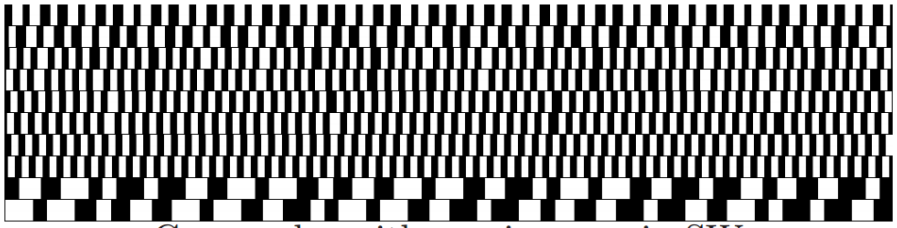
* indoor only

New Datasets

- 2011-2013 collected 33 new datasets
  - multi exposure, multi lighting
  - floating-point disparities (PFM)
- 2013-2014 improved processing at DLR
- Build on structured lighting method by Scharstein & Szeliski [CVPR 2003]

New structured lighting system 2010-2014

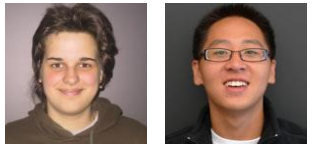
- Portable rig
  - 2 DSLRs, 2 consumer cameras
- Improved Gray codes
- Natural scenes
- Specular surfaces



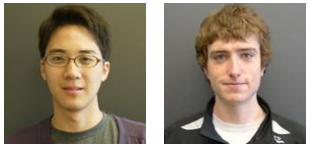
Gray codes with maximum min-SW

[Gupta et al., CVPR 2011]

2011: 5 datasets



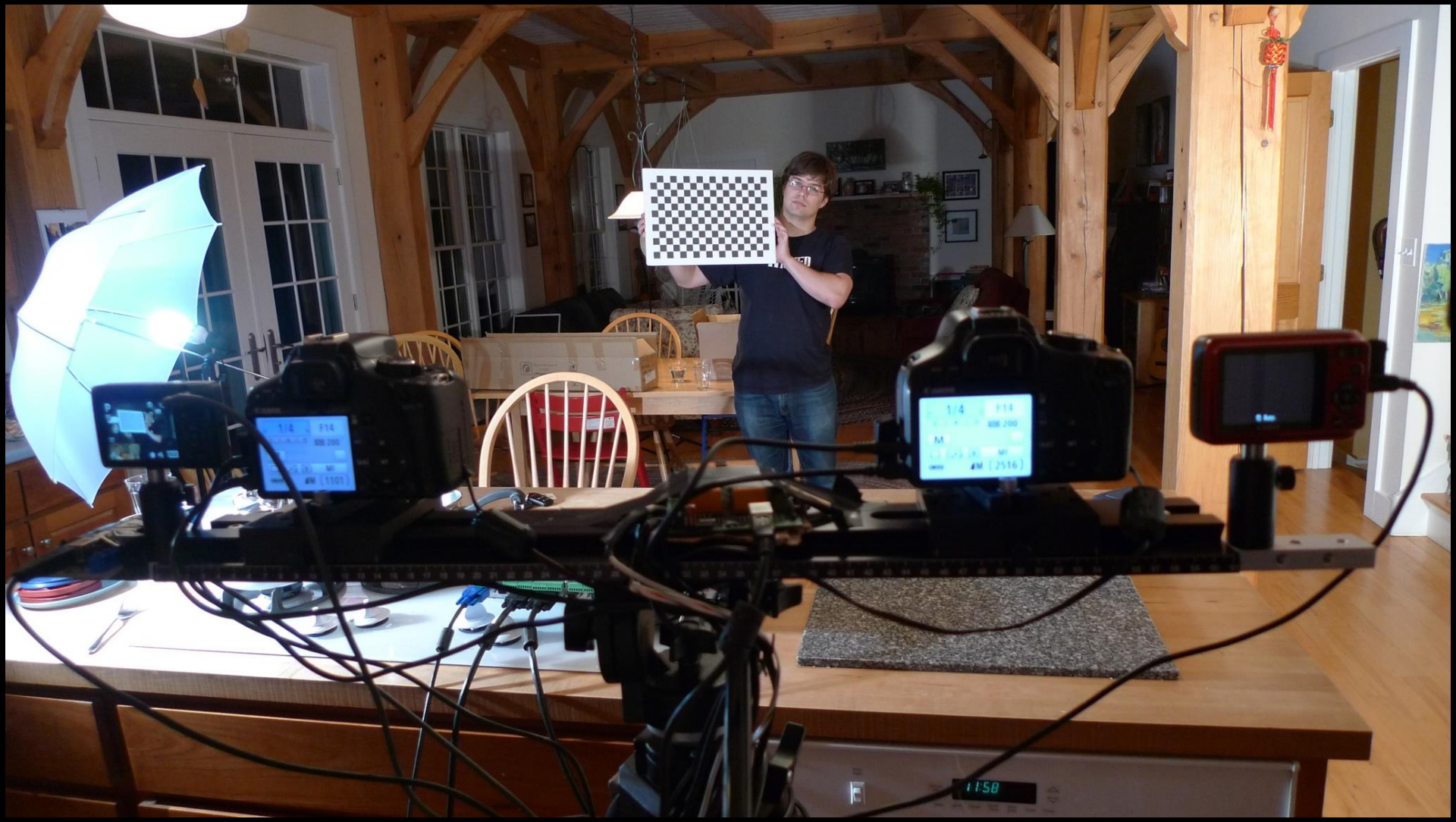
2012: 7 datasets



2013: 21 datasets

- 2014: Improved processing

2011



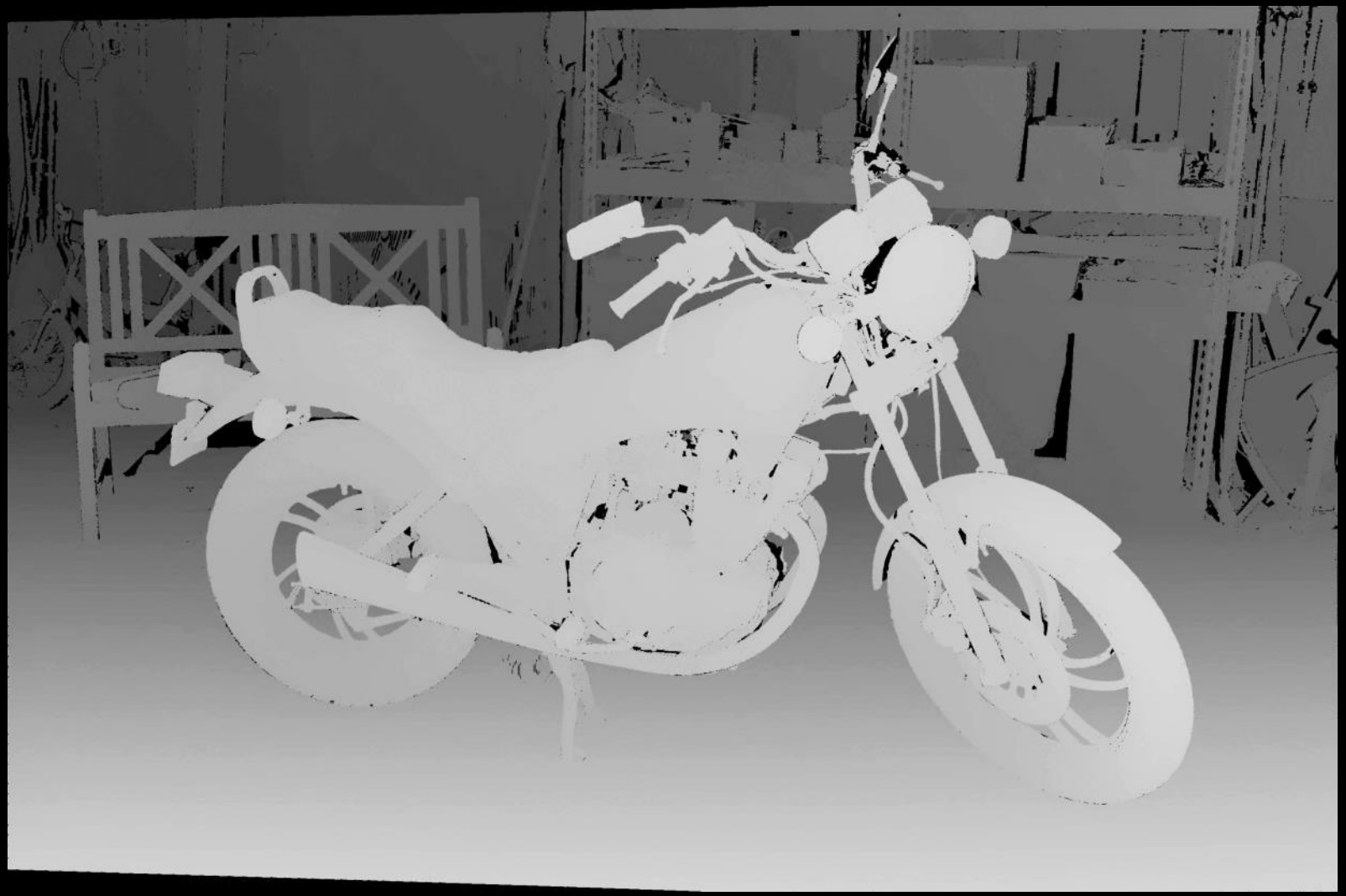

B

2012


YAMAHA

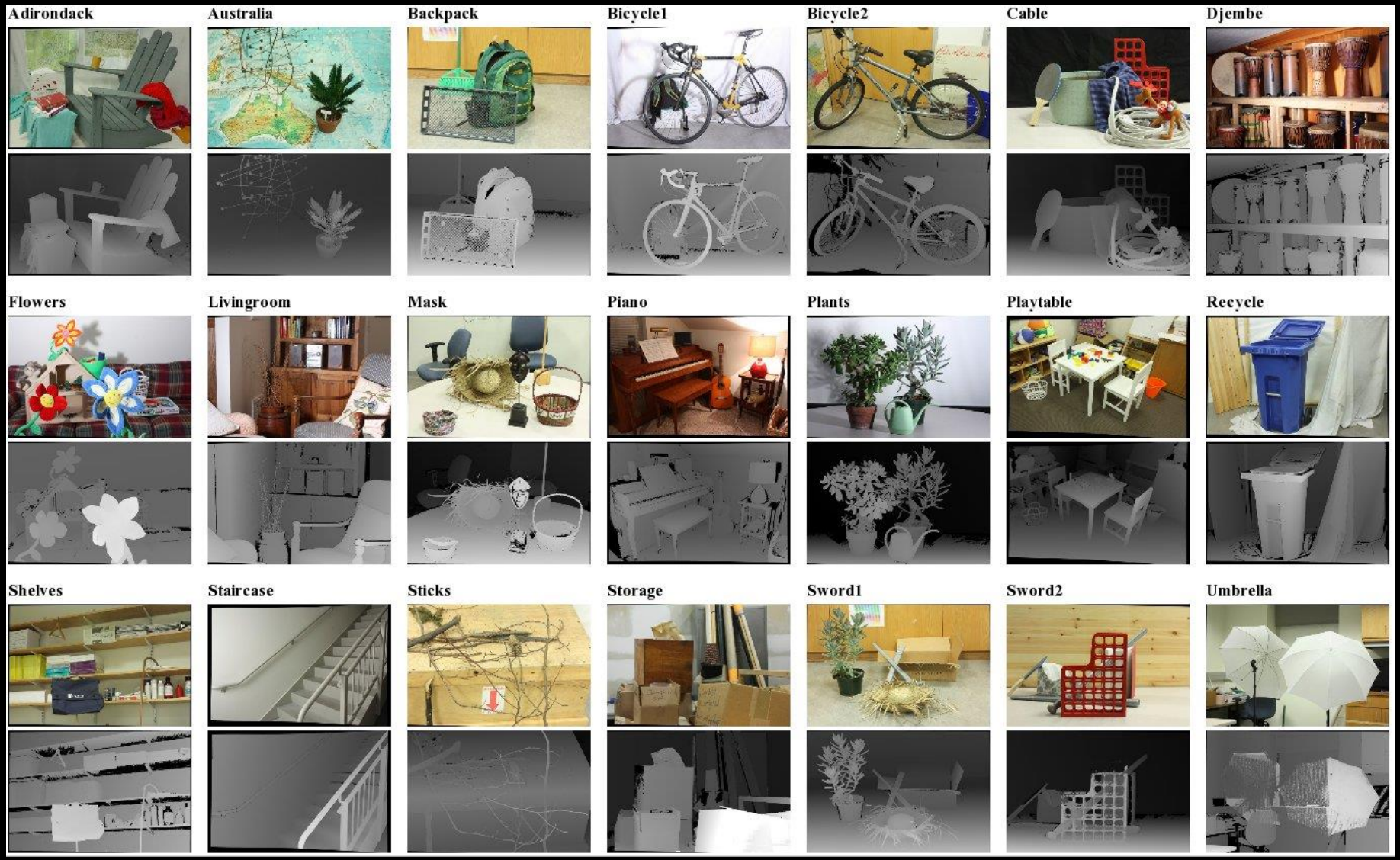

YAMAHA
RIDGID


YAMAHA
YAMAHA




YAMAHA

2013

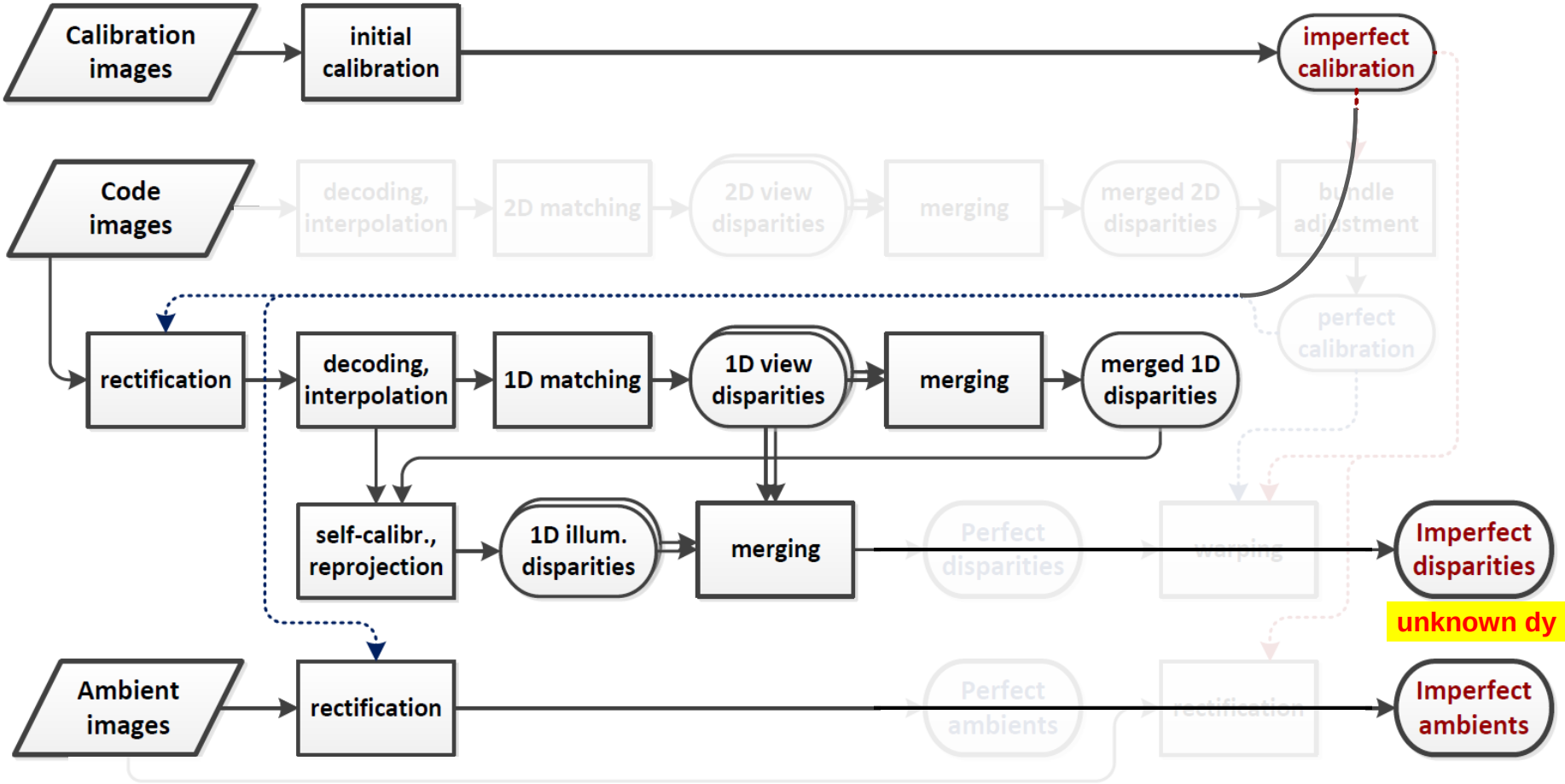

Adirondack
Australia
Backpack
Bicycle1
Bicycle2
Cable
Djembe
Flowers
Livingroom
Mask
Piano
Plants
Playtable
Recycle
Shelves
Staircase
Sticks
Storage
Sword1
Sword2
Umbrella

Improved processing

- How to get subpixel accuracy at 6 MP?
- Contributions:
  - Robust interpolation of codes
  - Fast 2D subpixel correspondence search
  - **Improved calibration via bundle adjustment**
  - **Improved self-calibration of projectors**
- Results:
  - Rectification and disparity accuracy of < 0.2 pixels
  - “perfect” and “imperfect” versions of datasets
- Best paper award at GCPR 2014

Processing pipeline

old system

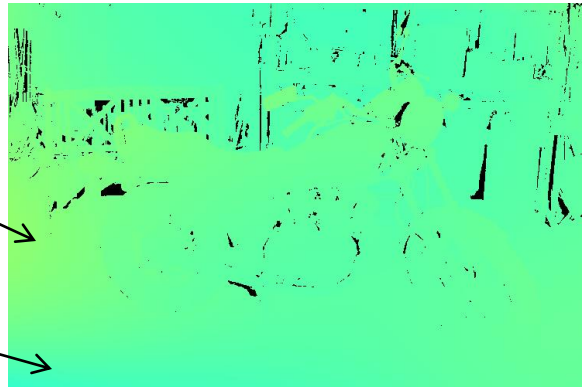


Rectification errors

y-disps

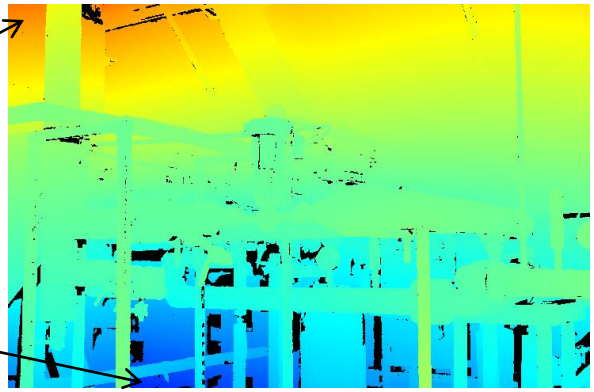
max dy = -0.25

min dy = -1.5



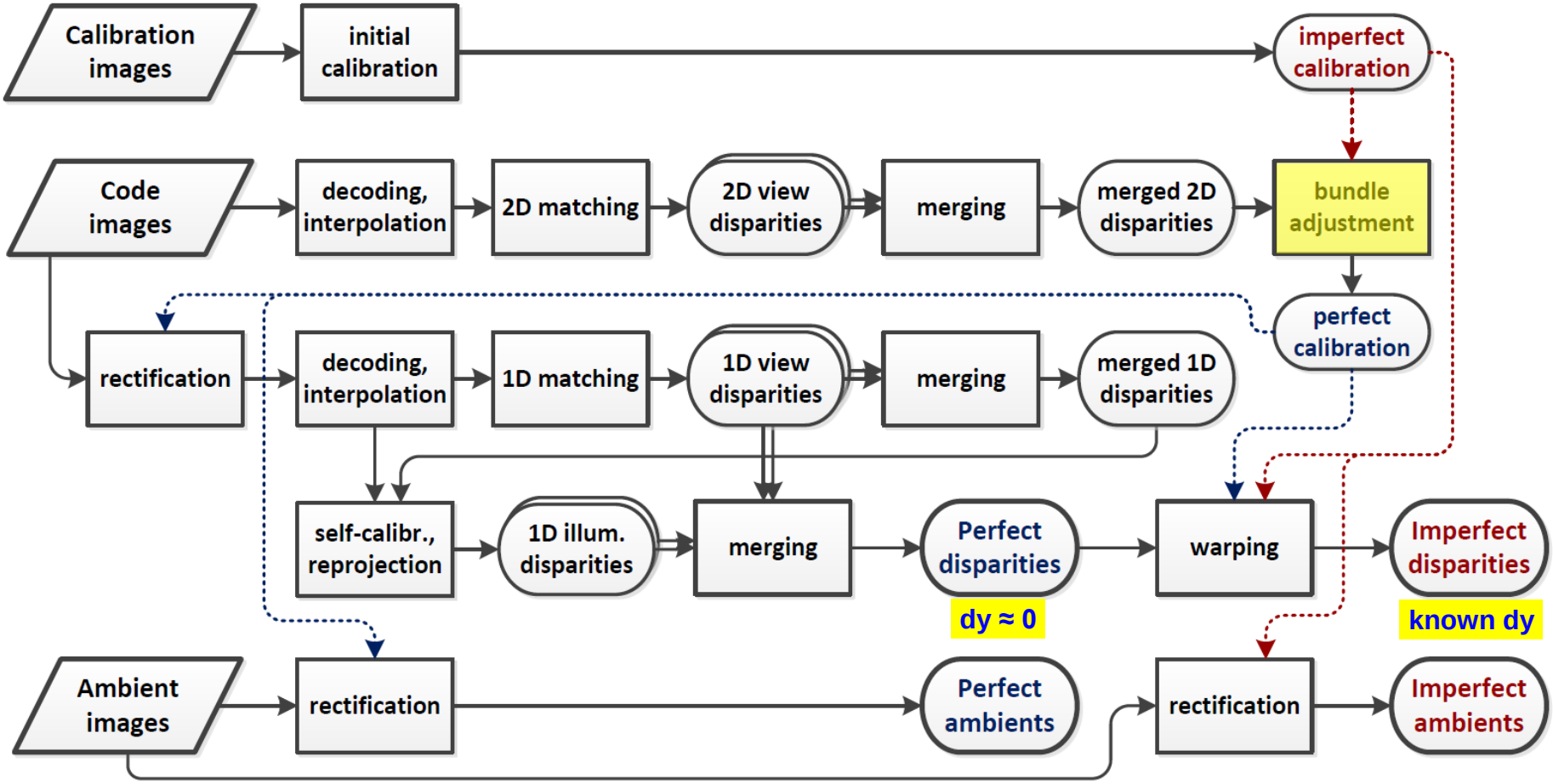
max dy = +3.2

min dy = -5.2



Processing pipeline

new system



Bundle adjustment

- Minimize residual y-disps using nonlinear opt.
- Refine subset of camera params:

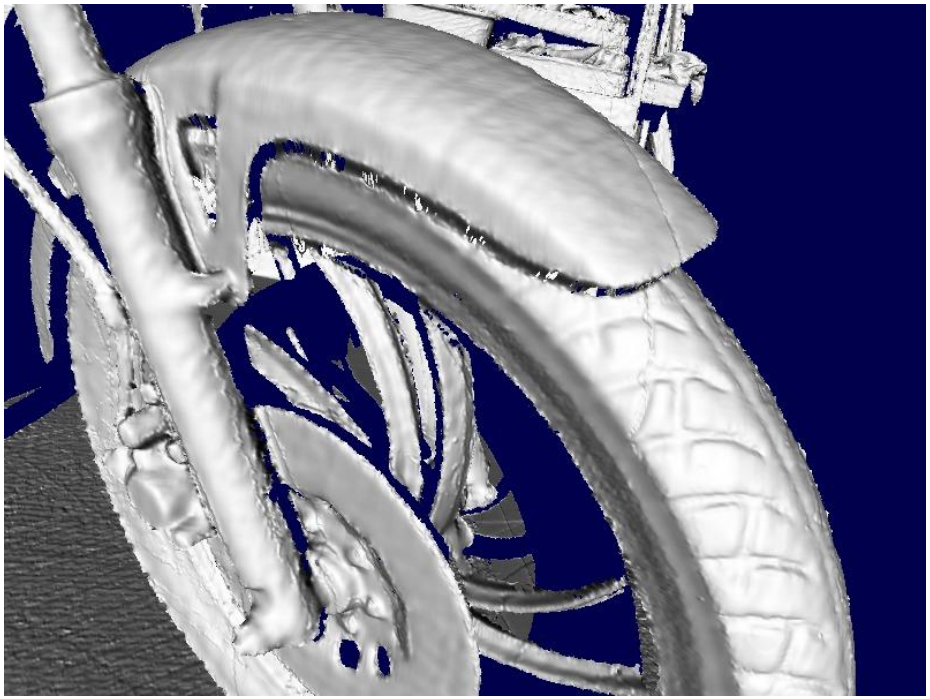
	left cam	right cam
lens distortion	κ_1 , κ_2	κ_1 , κ_2
intrinisics	f_x , f_y	$f_x = f_y$
	c_x , c_y	c_x , c_y
extrinisics	$\boldsymbol{R}_0$	$\boldsymbol{R}_1$
	$\boldsymbol{T}_0$	$\boldsymbol{T}_1$ fixing $\|\|\boldsymbol{T}_1 - \boldsymbol{T}_0\|\|$

Accuracy

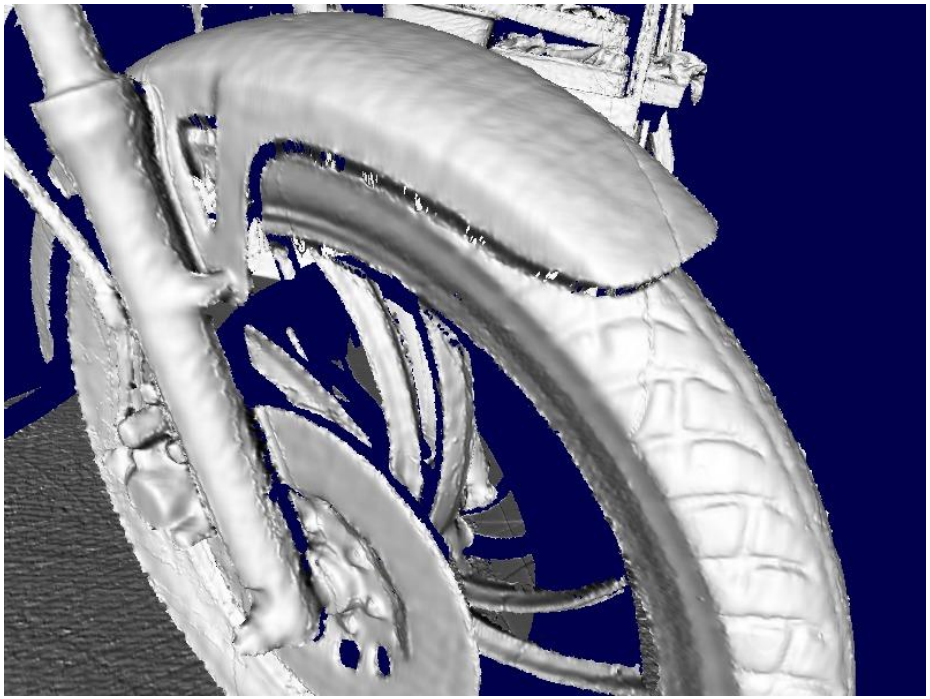
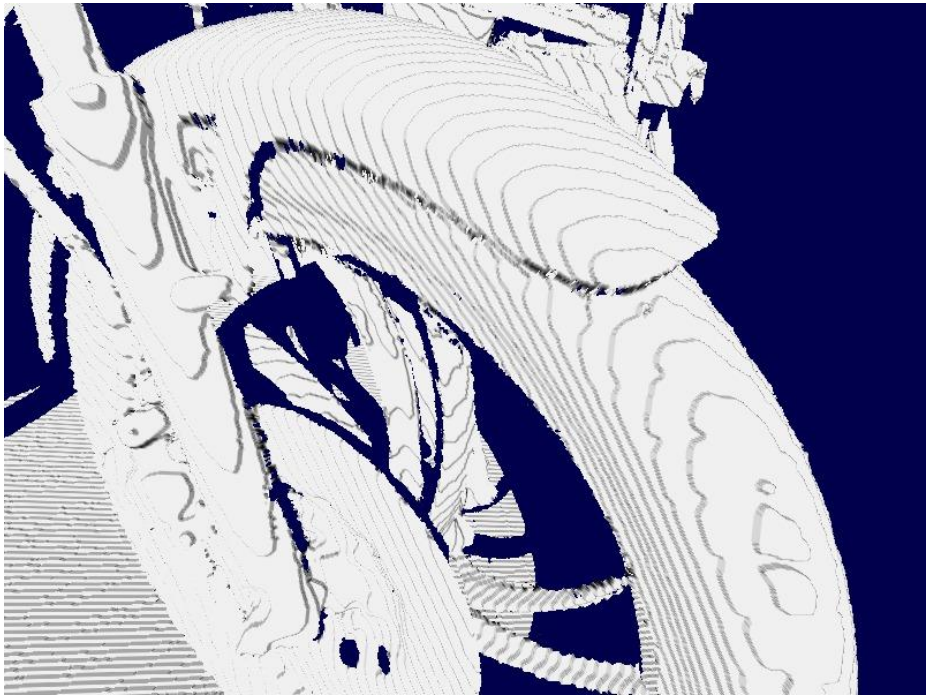
- No absolute measurements available, but
- Can check consistency of disparity estimates from P different projectors:
  - avg sample stddev $s = 0.20$
  - avg # of samples $n = 7.7$
  - we provide these as PFM images
- Can check residuals in planar regions

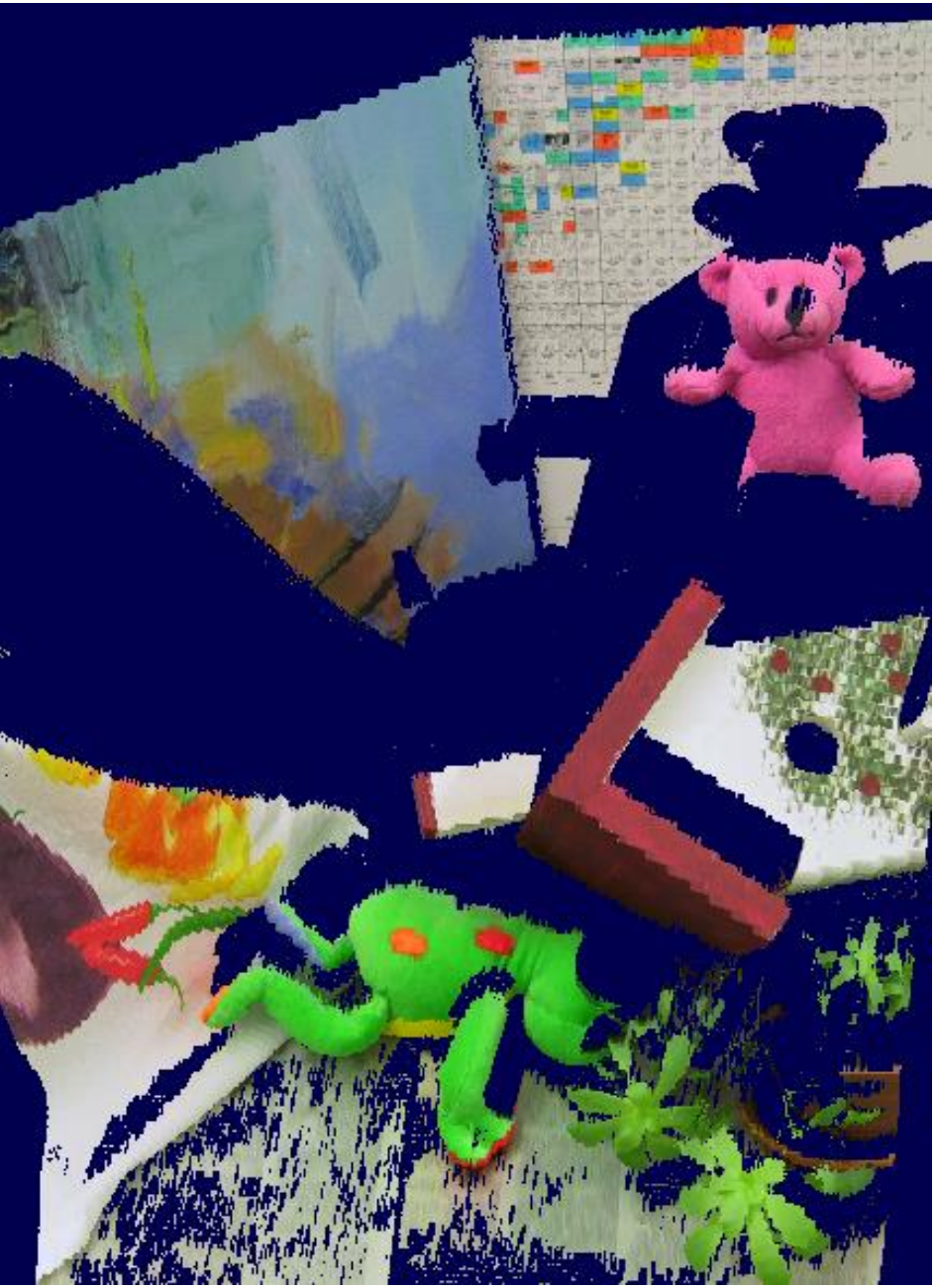
Average absolute residuals			Improvement	
int disps	no subpix	ours		
$\overline{r}_0$	$\overline{r}_1$	$\overline{r}_2$	$\overline{r}_0/\overline{r}_2$	$\overline{r}_1/\overline{r}_2$
0.252	**0.135**	**0.032**	**7.9**	**4.2**

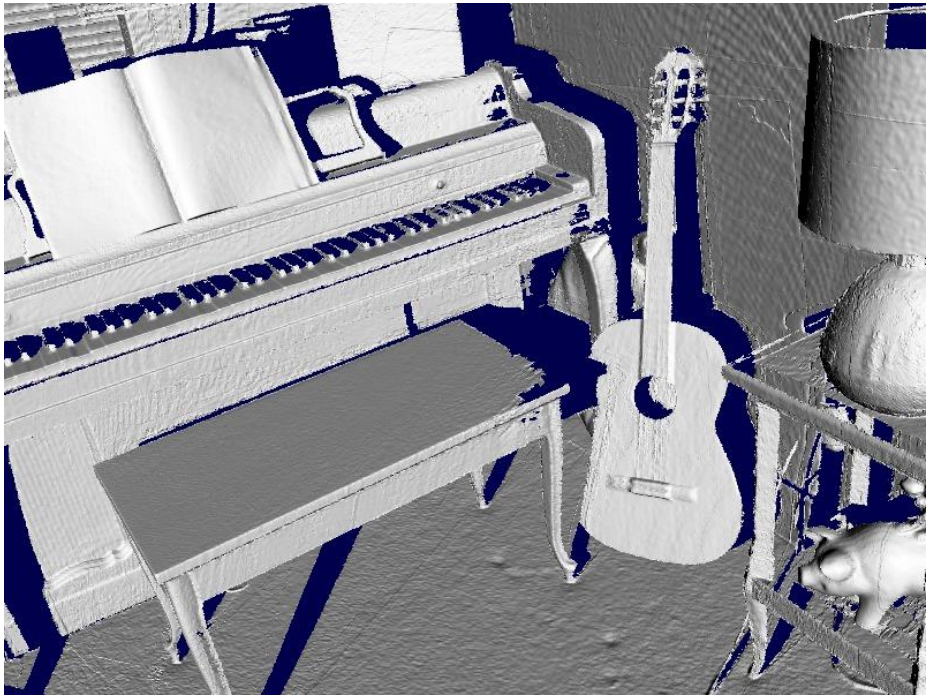


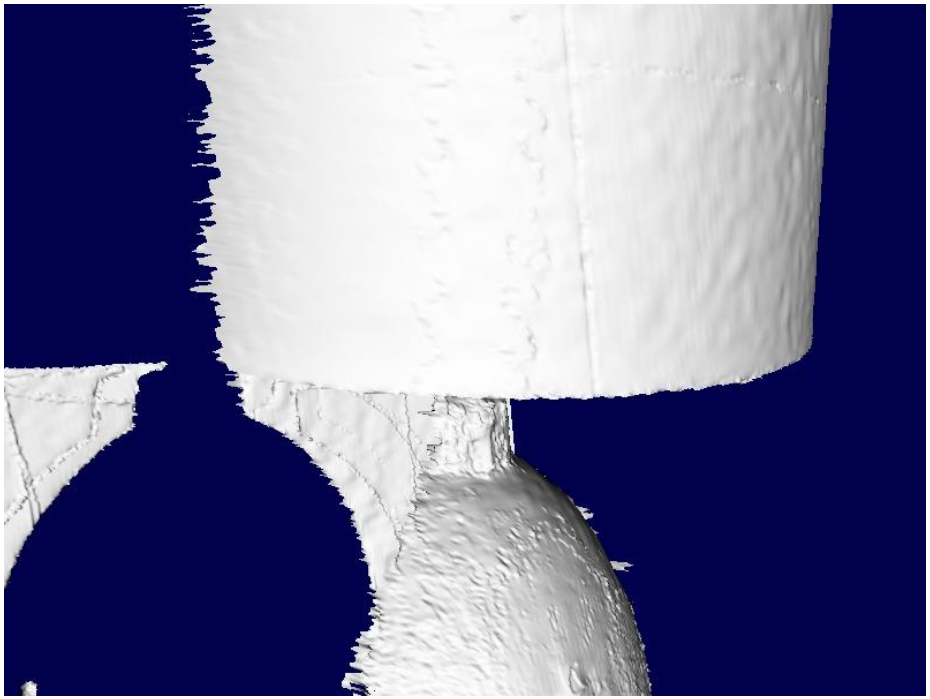
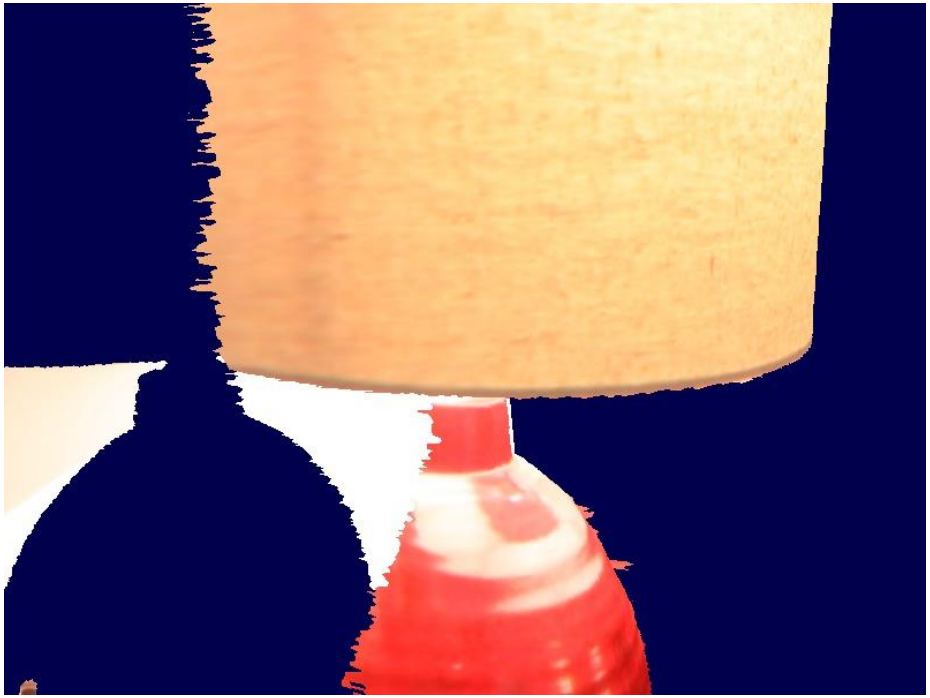


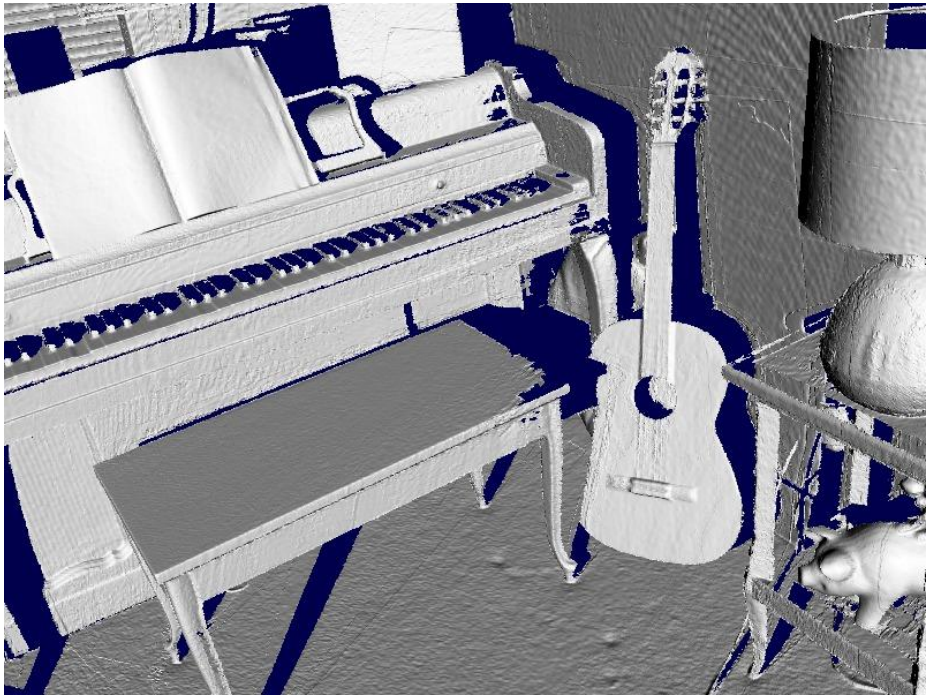


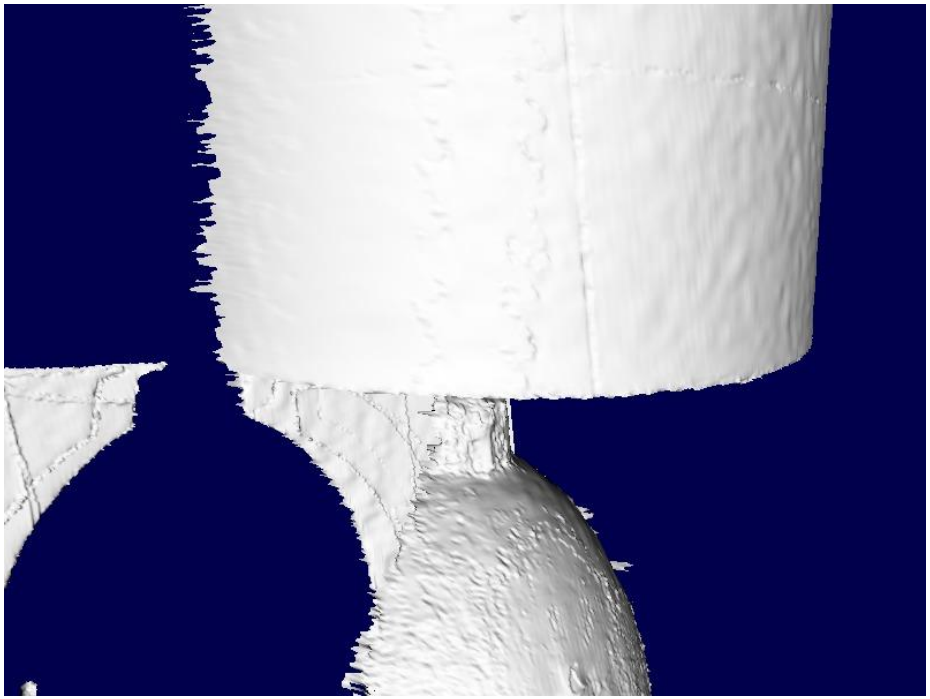
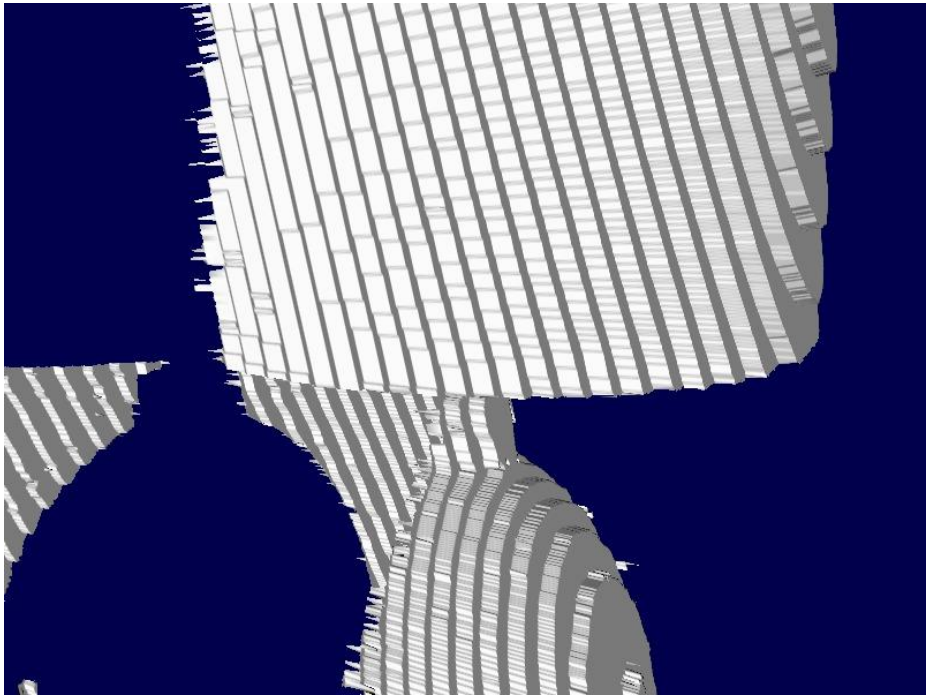






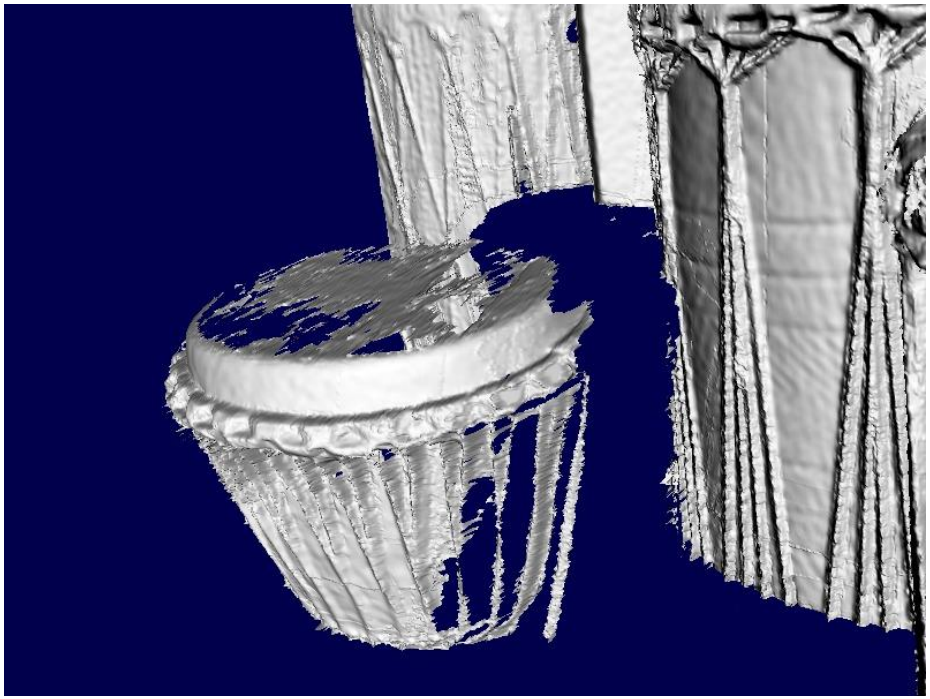


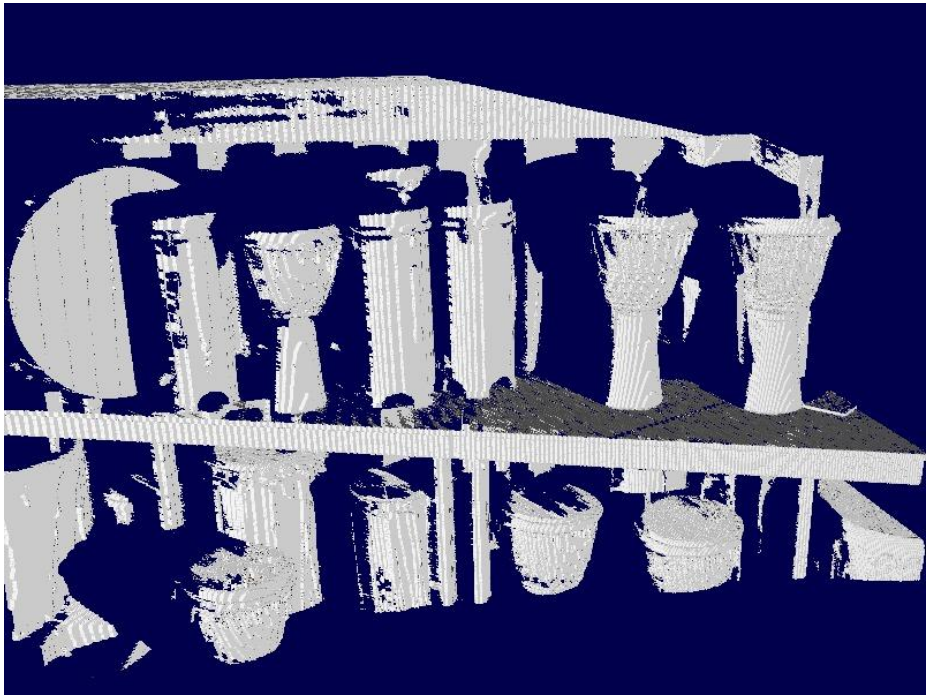
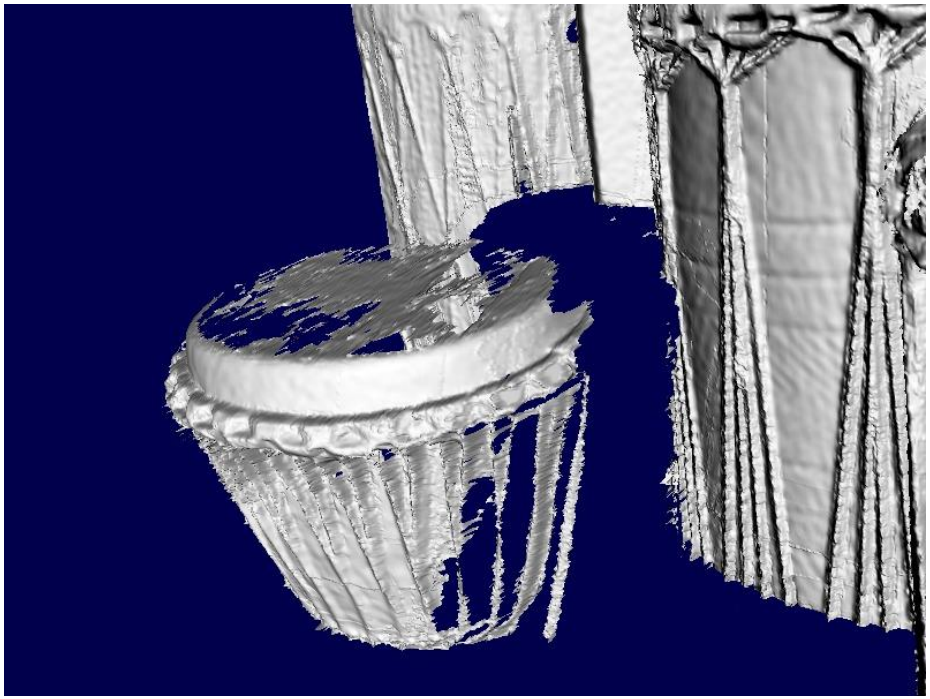
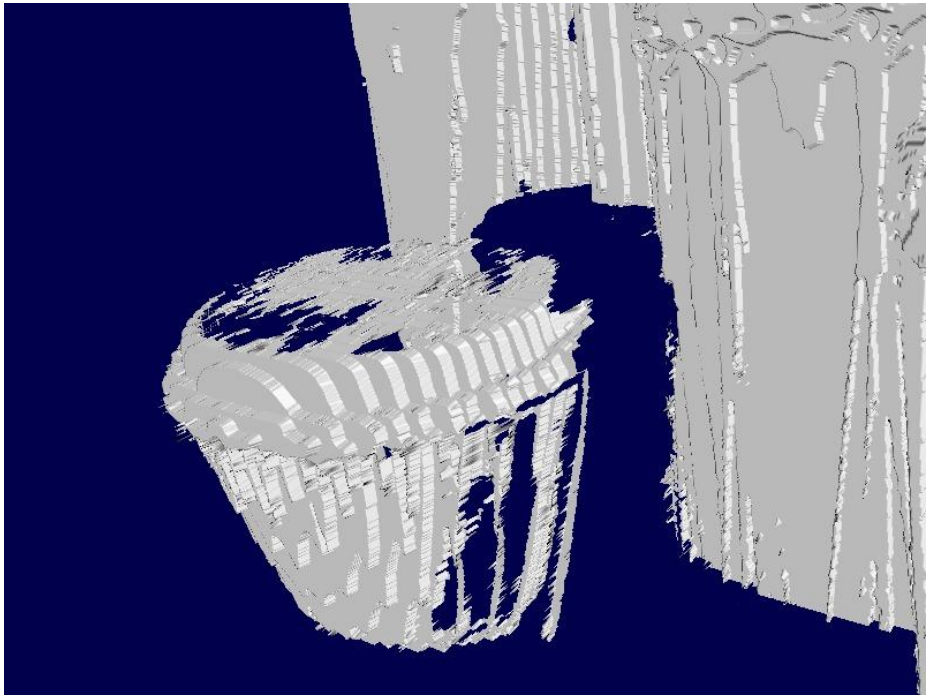








Demo sv / plyv (part of svkit software)

Benchmark Design – Goals

- Integrate lessons learned (Middlebury, KITTI, Sintel, …)
- Provide both overview and enable detailed interactive analysis
- Prevent overfitting; enable periodic re-calibration of difficulty

Middlebury Benchmark v.3

- Datasets:
  - 15 training pairs, 15 test pairs (hidden GT)
  - Full, Half, and Quarter resolution
  - Some pairs with varying exposure, illumination
  - Most pairs with imperfect rectification
- Evaluation:
  - Multiple performance metrics
  - Weighted average for overall ranking
  - Allows evaluation of “sparse” results

Metrics

- Bad pixel % (t=0.5, 1.0, 2.0, 4.0)
- Avg abs and RMS error
- Quantiles 50, 90, 95, 99
- Runtimes, also normalized by # of pixels (t/MP) and # of disp hypotheses (t/Gdisp)
- Evaluate both dense and sparse results
- Region masks: nonocc or all regions

Middlebury Benchmark v.3

- Web interface by Duncan Levear ’15
  - Automatic upload and evaluation
  - Interactive sorting and plotting
  - Selections persist between tables
  - Collect and display metadata
  - Supports history of snapshots
  - Supports periodic changes of weights

http://vision.middlebury.edu/stereo/eval3/

Overall Ranking

- Use weighted average instead of ranks
- Down-weight datasets too hard or too easy
- Allows adjusting of challenge as state of the art progresses
  - Adjust weights periodically
  - Could add more region masks in future
  - Could change default metric in future

Some personal insights

- “It’s the UI, stupid” ☺
- Huge value in compact representation and visualization of results
- Participation must be easy
- Strive for automatic scripts, but referee / moderator always needed
- Unless “one-shot contest,” can never completely avoid overfitting
- Significant time commitment

Wishlist

- Accurate GT for outdoor images
- “Internet vision” dataset w/ dense GT
- ~~Datasets for scene flow~~ [Menze & Geiger CVPR 2015]
- More efforts by others (Sintel & KITTI good start)
- More work on synthetic images

Conclusion

- Benchmarks are important, stimulate research
- Creating ground-truth data is challenging, fun
- Hi-resolution images require new level of accuracy
- Stereo is not a solved problem!