# Flexible priors for robust dense stereo matching

Sudipta N. Sinha

Microsoft Research

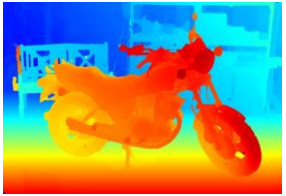
*Lines, Planes and Manhattan Models for 3-D Mapping Workshop, IROS 2017*

# Passive Stereo Matching

Much progress in 20+ years ..

[Okutomi et al. 1996]

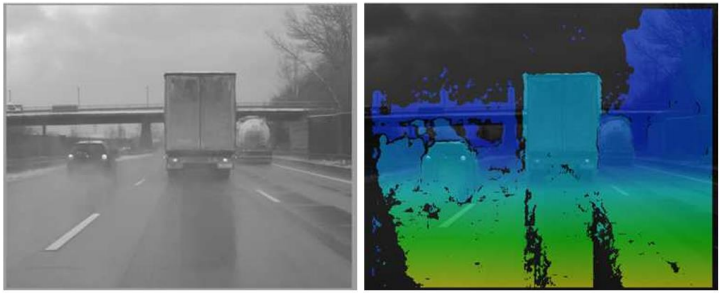
[Bleyer & Gelautz 2004]



[Taniai et al. 2017]

# Still challenging …

- Reflections, transparency
- Thin structures
- Untextured regions
- Outdoor weather
  - low light, rain, snow ..
- High resolution, real-time computation
- Imperfect calibration

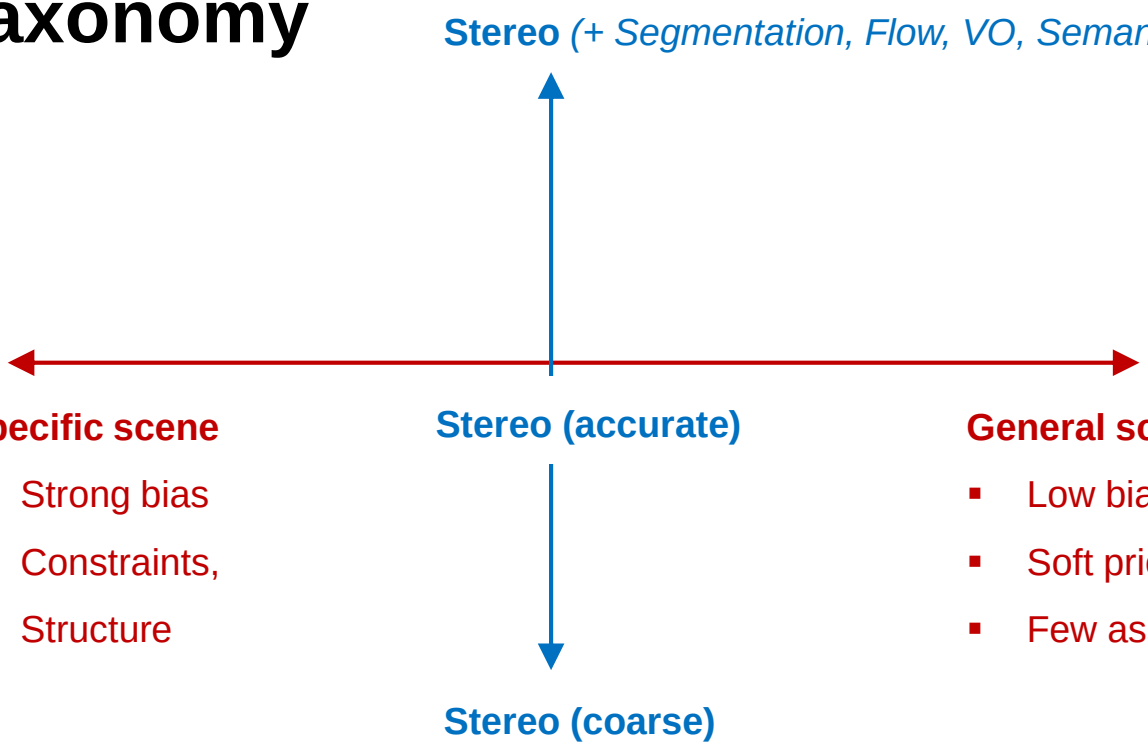


Nair+ ICCV 2015

# Incorporating priors and constraints

- Priors add robustness in presence of noise, ambiguity
- How to incorporate priors?
  - Traditional tools (MRF inference; energy minimization)
  - Solve stereo in conjunction with other tasks
  - Supervised learning (deep learning)
- Adding structure (or constraints) to model
- Priors must be adaptive, to be useful

# Taxonomy

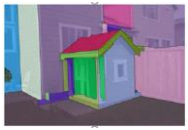

Stereo (+ Segmentation, Flow, VO, Semantics)
Specific scene
Strong bias
Constraints,
Structure
Stereo (accurate)
General scene
Low bias
Soft prior
Few assumptions
Stereo (coarse)

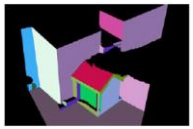
**Stereo** *(+ Segmentation, Flow, VO, Semantics)*

**Specific scene**

**Stereo (accurate)**

**General scene**



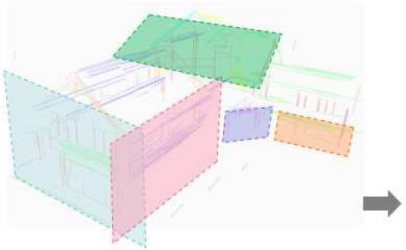


- Man-made scenes
- Image-based rendering

**Stereo (coarse)**

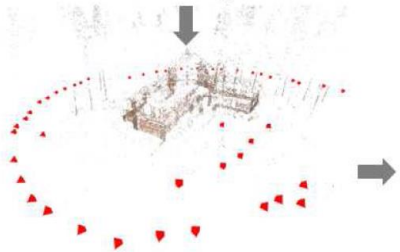
# Stereo matching with planar priors

[Sinha, Steedly and Szeliski 2009]

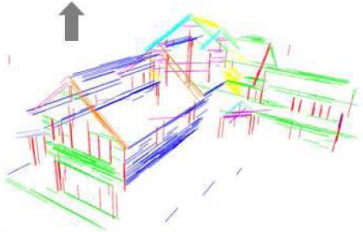


Multiple Plane Detection

MRF optimization

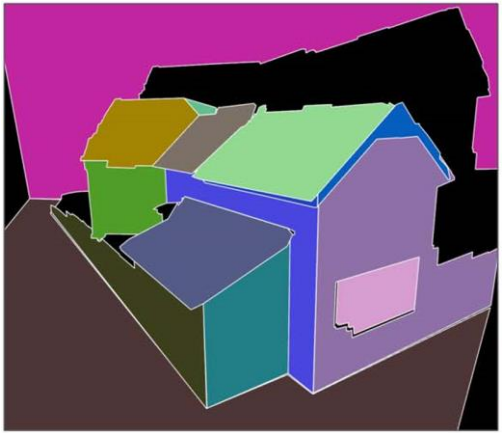


Structure from motion



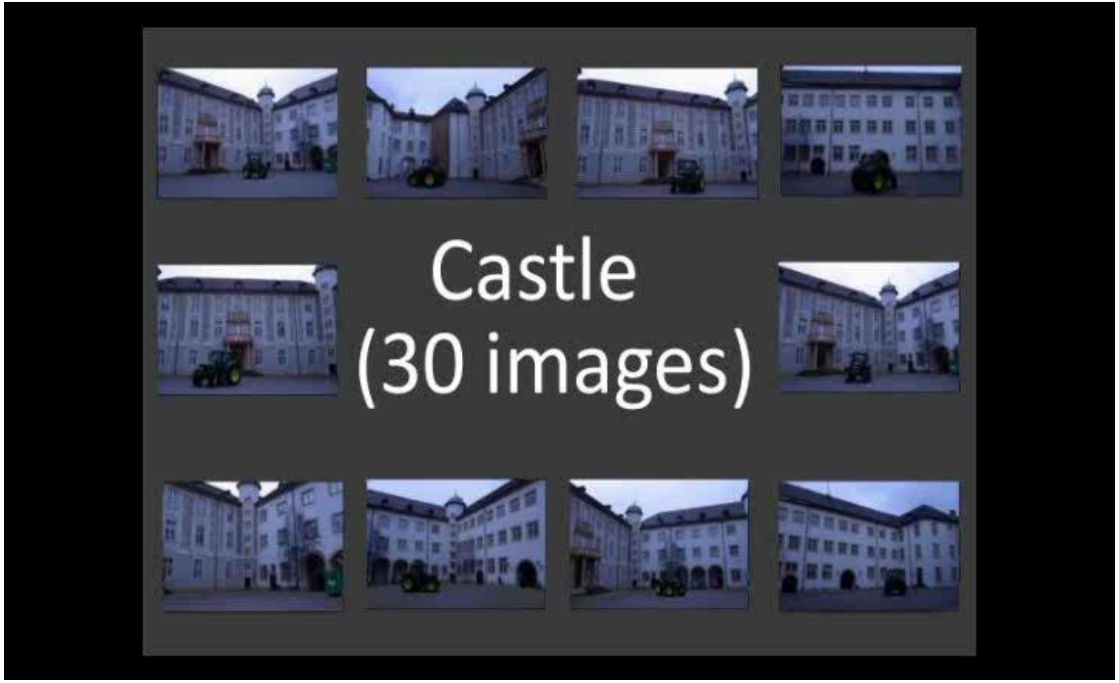
3D Line Reconstruction

# Stereo matching with planar priors



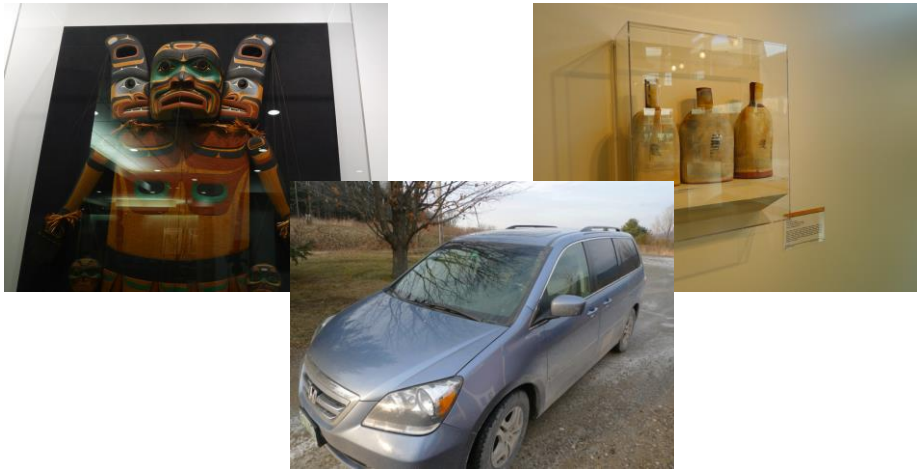
# IBR using piecewise planar depth maps

# IBR using piecewise planar depth maps



# Scenes with Reflections and Transparency

Reflective, glossy and transparent surfaces



# Scenes with Reflections and Transparency

[Sinha+ 2012]

- Model such scenes with two layers; reflections have depth
- Two-layer depth maps; piecewise planarity

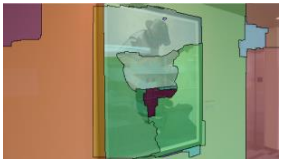
Input Image

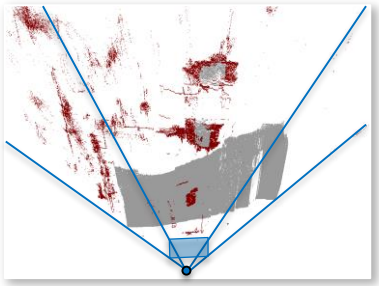
Depth Map (layer1)

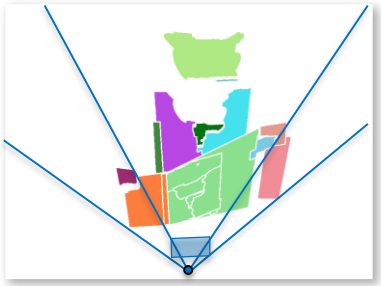
Depth Map (layer2)

# Scenes with Reflections and Transparency

- Decompose images (color) given the two depth layers





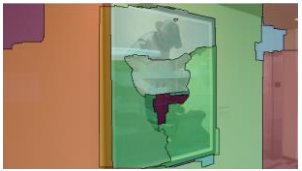


Two-layer stereo matching

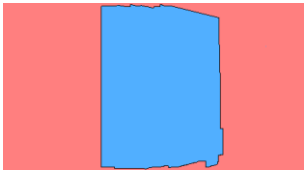
# Scenes with Reflections and Transparency

# Scenes with Reflections and Transparency

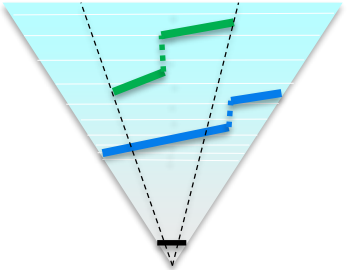
- **Two-layer Piecewise Planar Stereo**
  - MRF Labels:
    - Single Planes **(opaque)**
    - Pair of Planes **(opaque + reflective)**
  - Energy minimization via graph cuts



Labeling

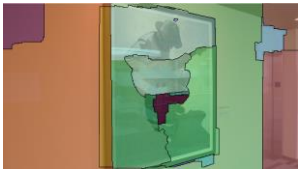


Reflectivity Map

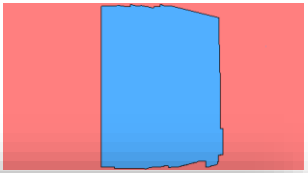


# Scenes with Reflections and Transparency

- **Two-layer Piecewise Planar Stereo**
  - MRF Labels:
    - Single Planes **(opaque)**
    - Pair of Planes **(opaque + reflective)**
  - Energy minimization via graph cuts



Labeling



Reflectivity Map

- BUT, recovered reflectivity mask often inaccurate
- Still, many open issues in reflection modeling

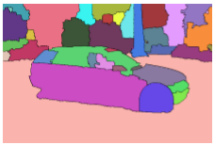
**Stereo** *(+ Segmentation)* *Flow, VO, Semantics)*

Specific scene

Stereo (accurate)

**General scene**

**Stereo (coarse)**

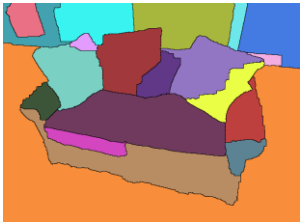


- Image-based rendering
- Image editing

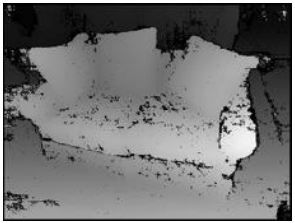
# Stereo matching with planar priors

[Kowdle, Sinha and Szeliski 2012]

- Tackle general scenes
- Local planes as primitives
- Estimate pixel-to-plane labeling
- Learn surface appearance (color) models

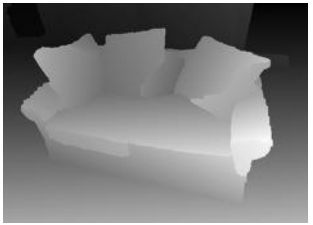


Plane label map



Semi-global stereo (SGM)

Find planes



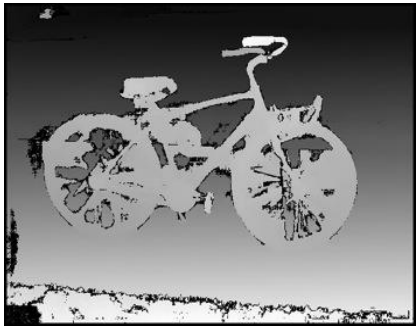
Depth map

# Stereo matching with planar priors

[Kowdle, Sinha and Szeliski 2012]

Iterate between:

- Segmentation (graph cuts)
- Learning color distributions (like GrabCut [Rother+ 2004])



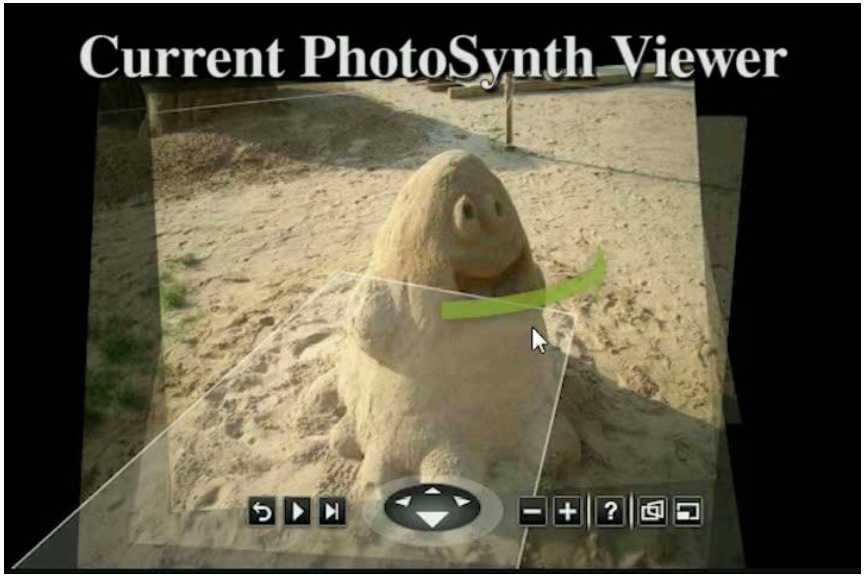
SGM stereo

Only stereo cues

Appearance + stereo cues

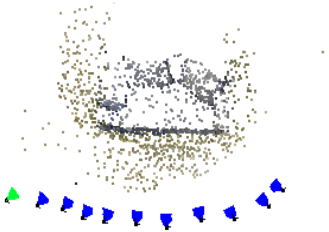
# Microsoft Photosynth2 (2012)

- Interactive, 3D immersive viewer for photos
- MSR SpinMovies prototype



# Automatic Multi-view Object Segmentation

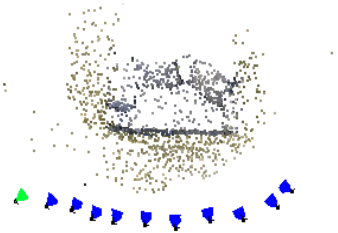
[Kowdle, Sinha and Szeliski 2012]



structure from motion (SfM) reconstruction

# Automatic Multi-view Object Segmentation

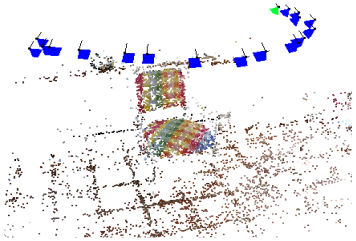
[Kowdle, Sinha and Szeliski 2012]



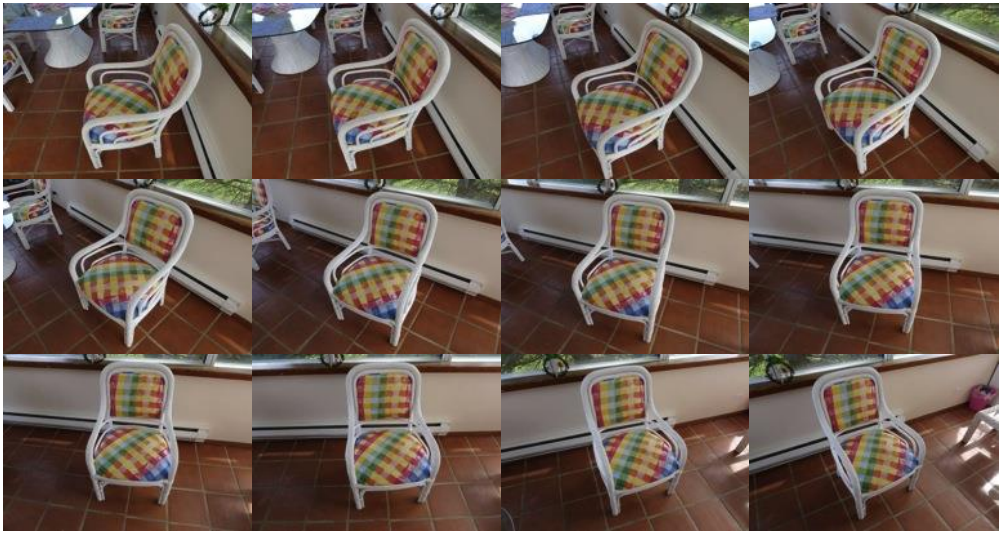
structure from motion (SfM) reconstruction

# Automatic Multi-view Object Segmentation

[Kowdle, Sinha and Szeliski 2012]

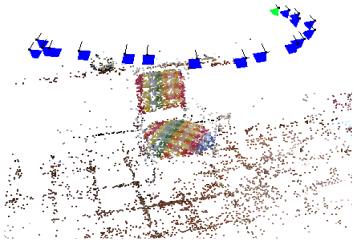


structure from motion (SfM) reconstruction

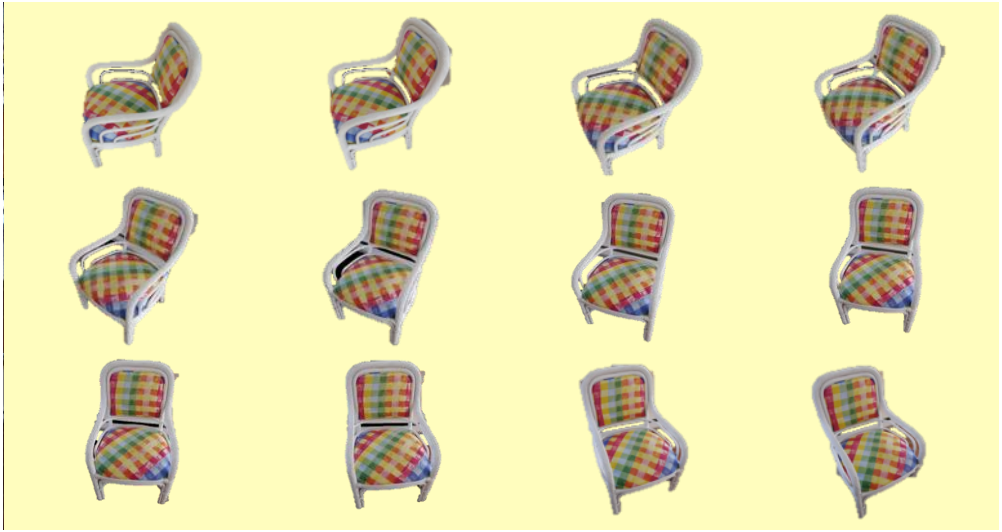


# Automatic Multi-view Object Segmentation

[Kowdle, Sinha and Szeliski 2012]

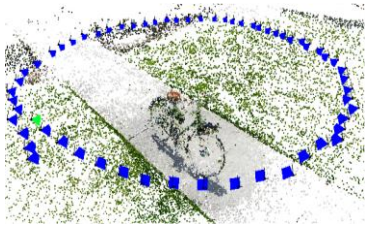


structure from motion (SfM) reconstruction

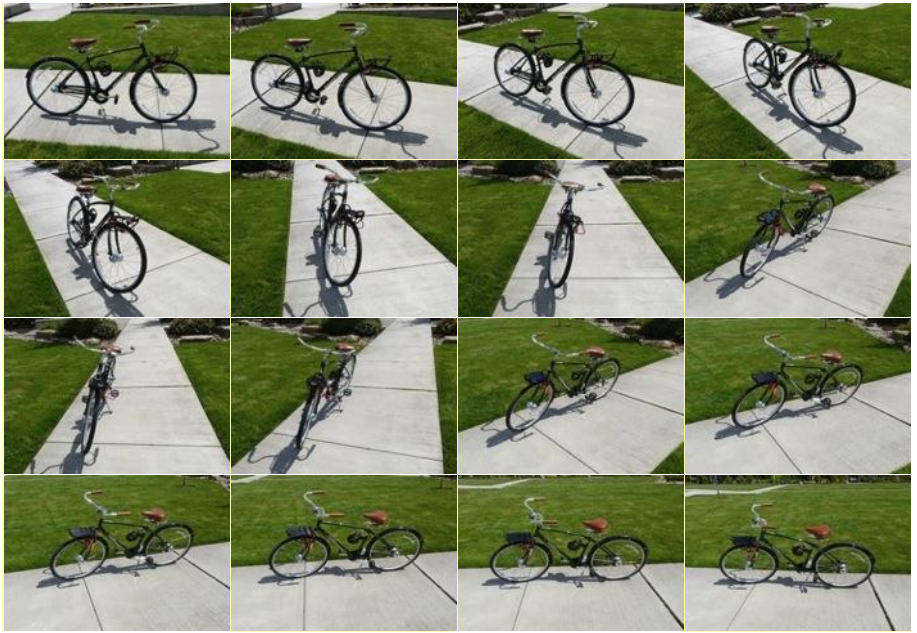


# Automatic Multi-view Object Segmentation

[Kowdle, Sinha and Szeliski 2012]

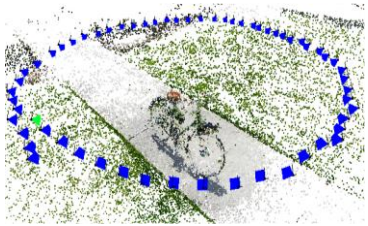


structure from motion (SfM) reconstruction



# Automatic Multi-view Object Segmentation

[Kowdle, Sinha and Szeliski 2012]



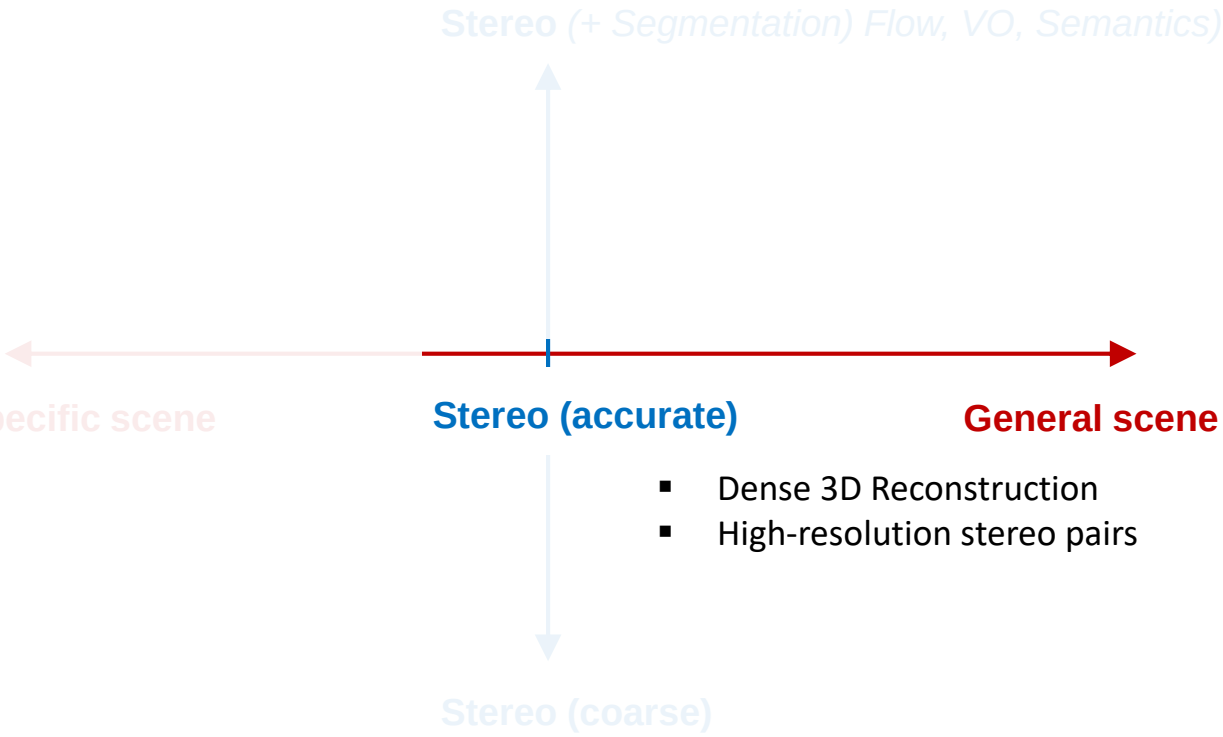
structure from motion (SfM) reconstruction

# Piecewise planar priors / IBR

- Accurate depth discontinuities and surface normals
  - more critical than accurate depth
- Planes (surfaces) as MRF labels instead of disparities.
  - Strong prior
  - Recovering planes (various ways)
- Restricted label set enforces geometric constraints
  - planarity, parallelism, two-layers ..

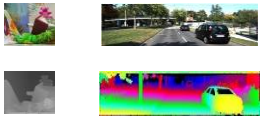
# Piecewise planar priors / IBR

- Limitations of strong priors
  - Strong bias; Underlying 3d shape lacks detail
- Common cause of failure
  - Proposal generator could miss some surfaces!
- Data-augmentation for deep learning
  - Currently based on image processing + CG rendering
  - Image-based rendering could also be effective !



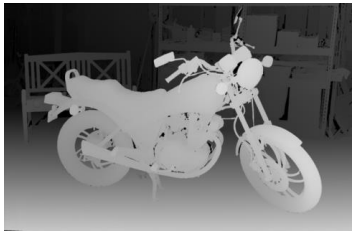

- Dense 3D Reconstruction
- High-resolution stereo pairs

Middlebury
Teddy (2005)



KITTI
(2012)

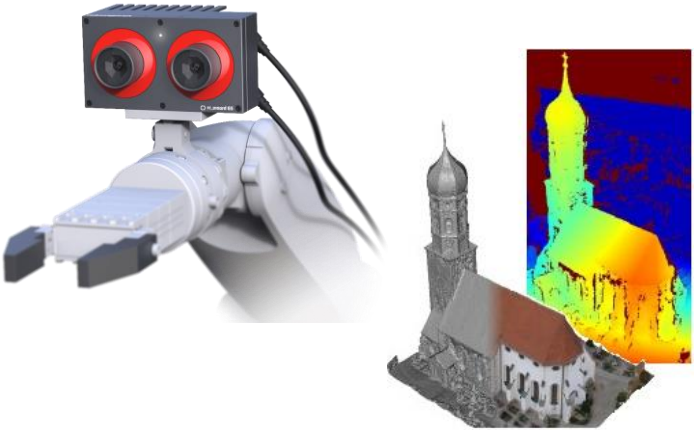
Middlebury v3 (2016)
5-6 Mpixels.



Disney (Kim+ 2013)
10—20 Mpixels.

# Semi Global Matching [Hirschmüller 2005]

- Most successful practical stereo method
- Used in photogrammetry, assisted driving, robotics, …

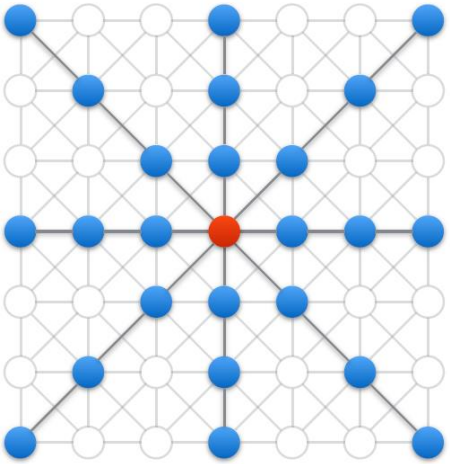



# Semi Global Matching [Hirschmüller 2005]

Approximates 2D MRF using 1D optimization

along 8 cardinal directions

$$E(D) = \sum_{\mathbf{p}} C_{\mathbf{p}}(d_{\mathbf{p}}) + \sum_{\mathbf{p},\mathbf{q}\in\mathcal{N}} V(d_{\mathbf{p}}, d_{\mathbf{q}})$$



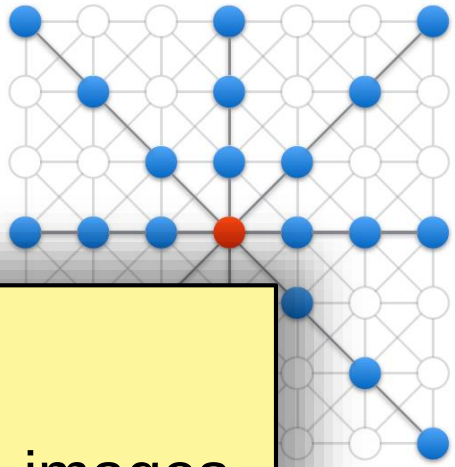
- related to BP, TRW-S [Drory et al. 2014]

# Semi Global Matching [Hirschmüller 2005]

Approximates 2D MRF using 1D optimization

along 8 cardinal directions

$$E(D) = \sum_{\mathbf{p}} C_{\mathbf{p}}(d_{\mathbf{p}}) + \sum_{\mathbf{p},\mathbf{q}\in\mathcal{N}} V(d_{\mathbf{p}}, d_{\mathbf{q}})$$



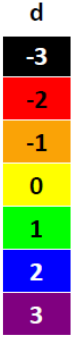
- rela

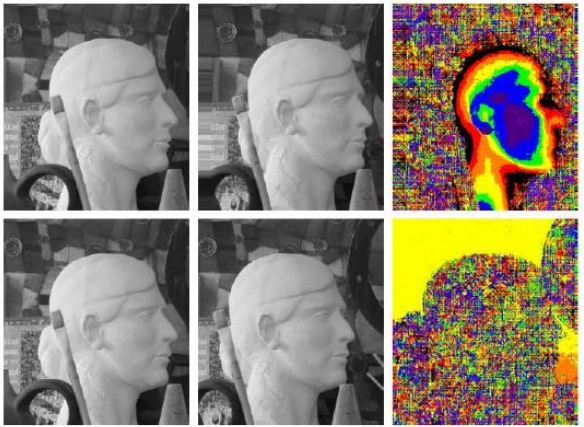
- Evaluates the whole DSI
- Inefficient for high-resolution images

# Local Plane Sweep (LPS) Stereo

[Sinha+ 2014]

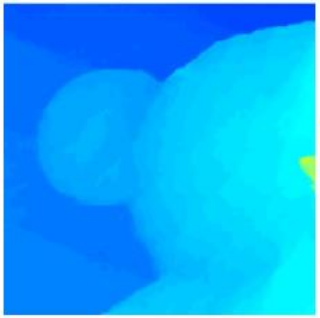
- Solve many local plane sweep stereo (LPS) problems
- Generates *surface* proposals*;* fuse them into a disparity map

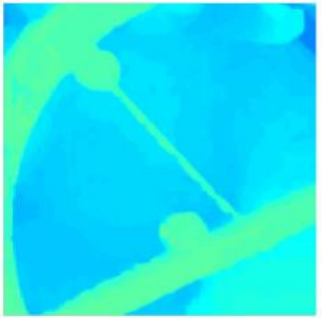



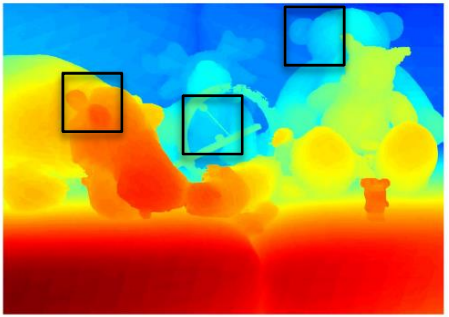



# Local Plane Sweep (LPS) Stereo

[Sinha+ 2014]

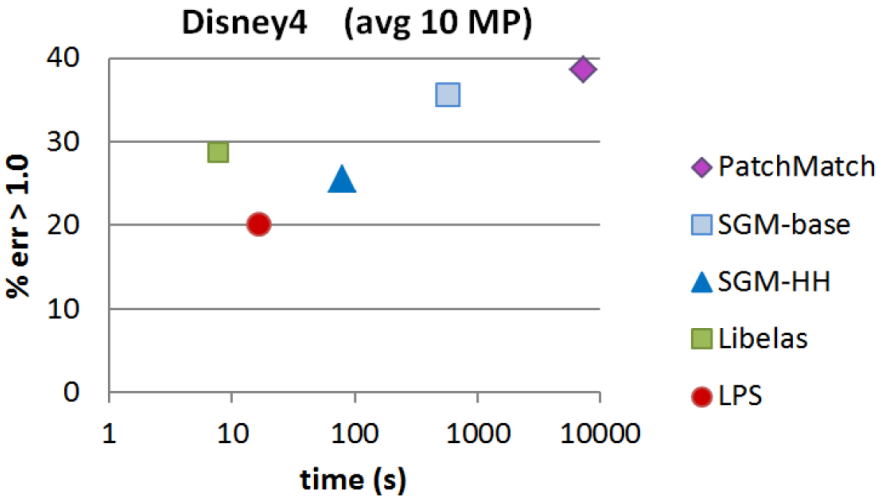


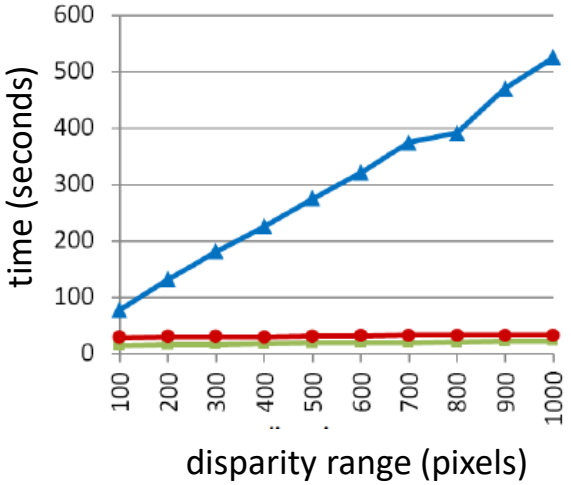




# Local Plane Sweep (LPS) Stereo

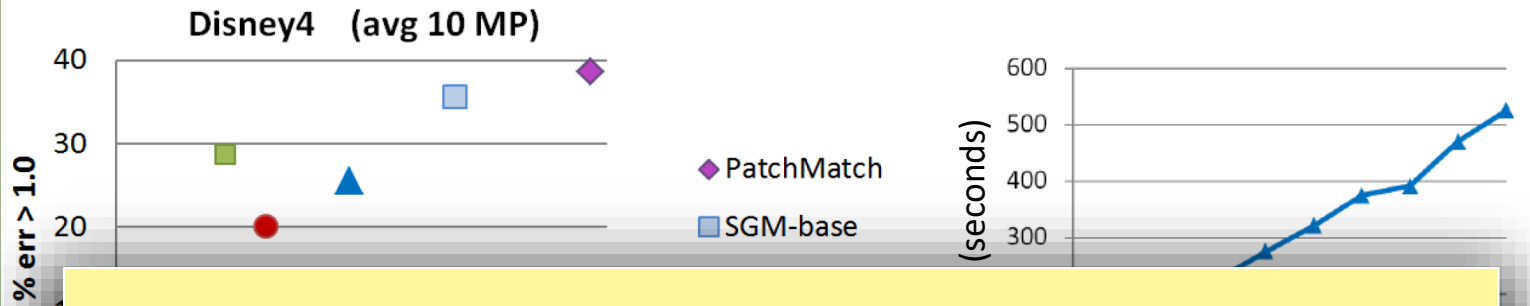
[Sinha+ 2014]



# Local Plane Sweep (LPS) Stereo

[Sinha+ 2014]




- more accurate than SGM; faster
- 2nd fastest after ELAS (Geiger+2010)
- BUT, plane fitting can be unreliable; can give errors

# Semi Global Matching [Hirschmüller 2005]

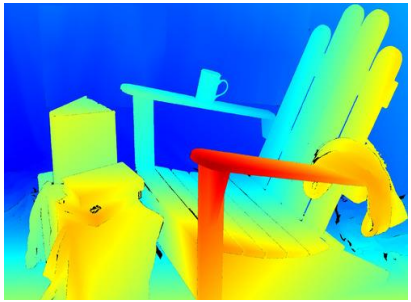
Approximates 2D MRF using 1D optimization along 8 cardinal directions

$$E(D) = \sum_{\mathbf{p}} C_{\mathbf{p}}(d_{\mathbf{p}}) + \sum_{\mathbf{p},\mathbf{q}\in\mathcal{N}} V(d_{\mathbf{p}}, d_{\mathbf{q}})$$

$$\begin{cases} 0 & \text{if } d = d' \end{cases}$$

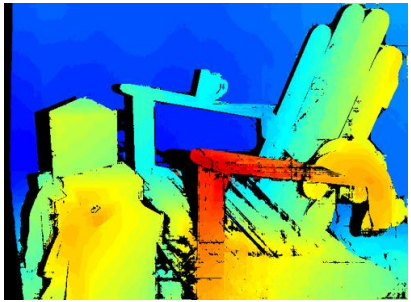
- Fronto parallel bias
- Inaccurate on slanted untextured surfaces

Left image

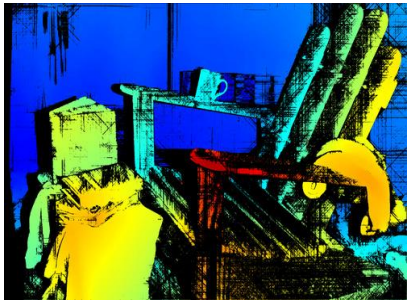


GT disparities

SGM @ quarter resolution



SGM @ full resolution (6 MP)



# SGM-P: SGM with orientation priors

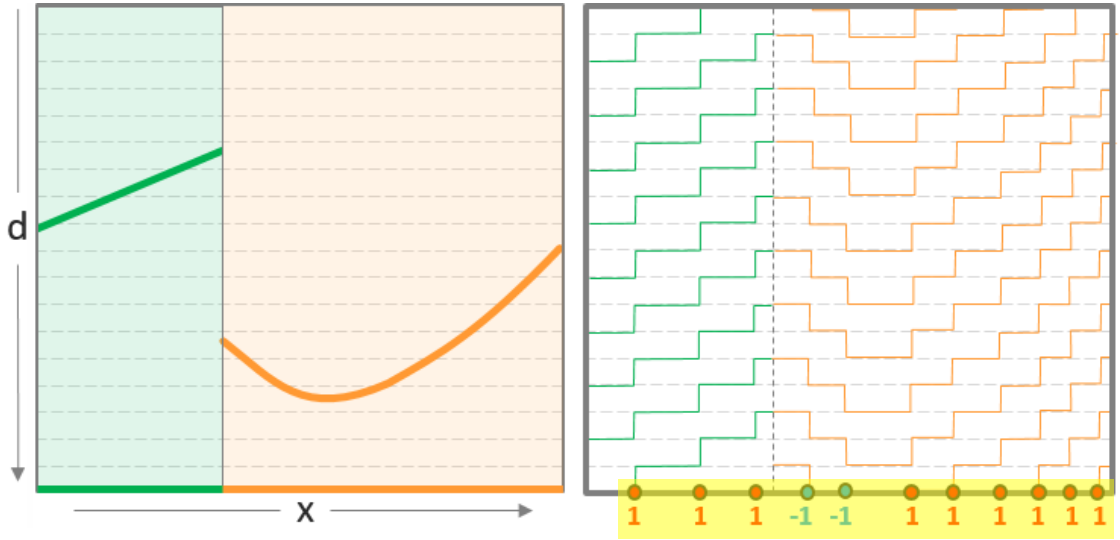
[Scharstein, Taniai, Sinha, 3DV 2017]

- What if we knew the surface slant?
- Replace fronto-parallel bias with bias parallel to surface

Idea:

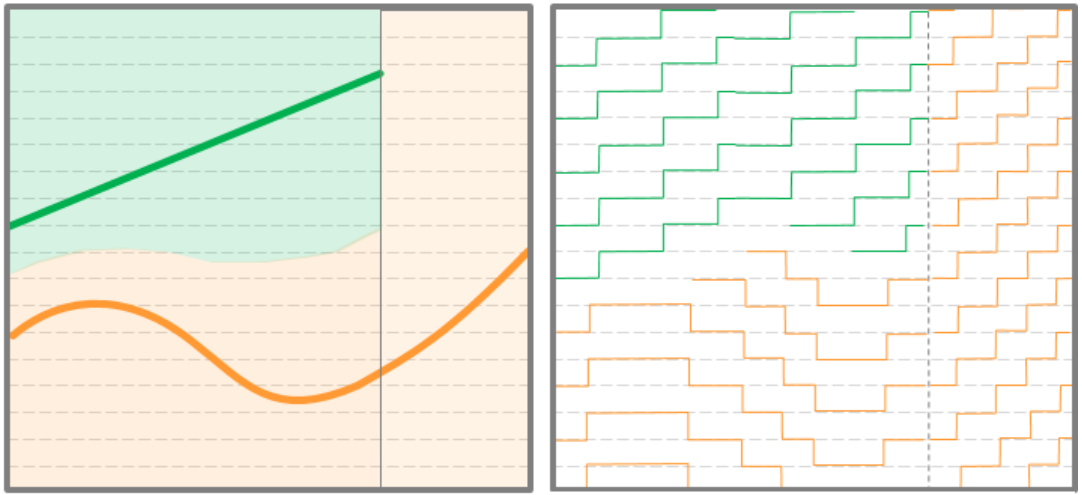
- *Rasterize* disparity surface prior (at arbitrary depth)
- Adjust $V(d, d')$ to follow discrete disparity "steps"

# SGM-P: 2D orientation priors




$$V_S(d_{\mathbf{p}}, d'_{\mathbf{p}}) = V(d_{\mathbf{p}} + j_{\mathbf{p}}, d'_{\mathbf{p}})$$

# SGM-P: 3D orientation priors



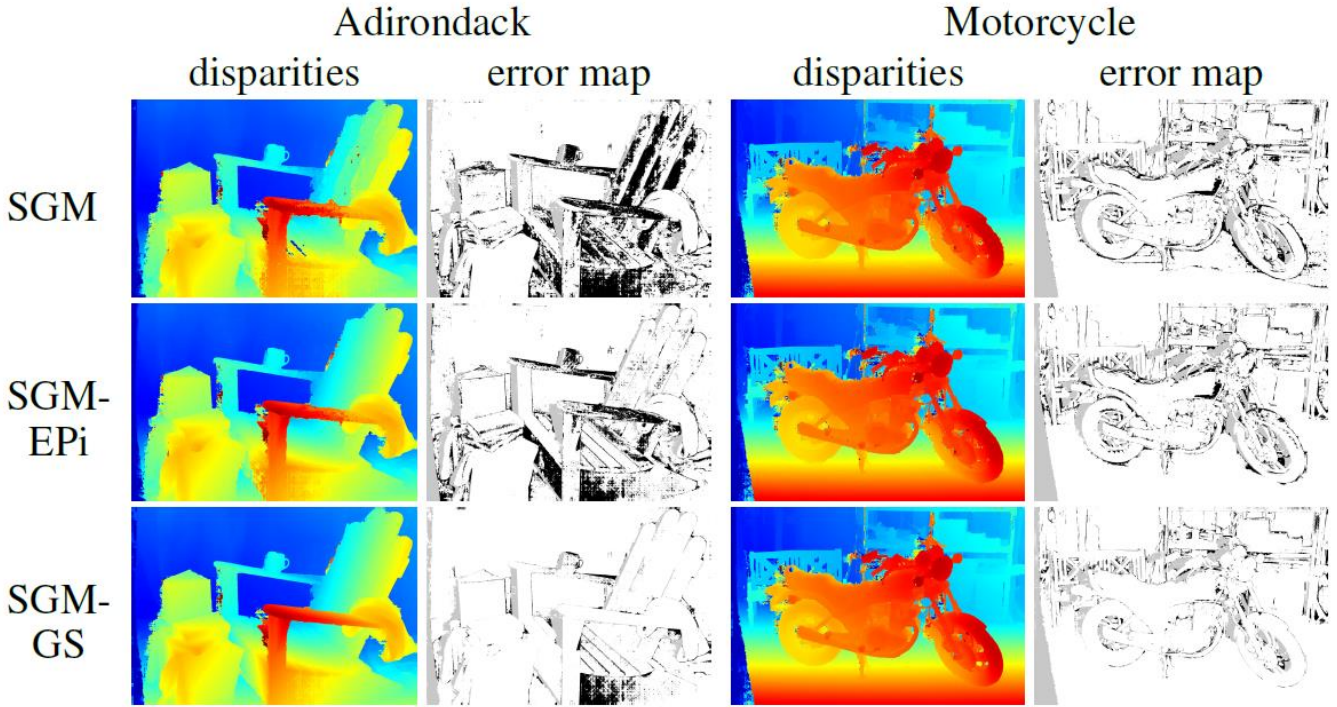
$$V_S(d_{\mathbf{p}}, d'_{\mathbf{p}}) = V(d_{\mathbf{p}} + j_{\mathbf{p}}(d_{\mathbf{p}}), d'_{\mathbf{p}})$$

Jump locations vary with disparity

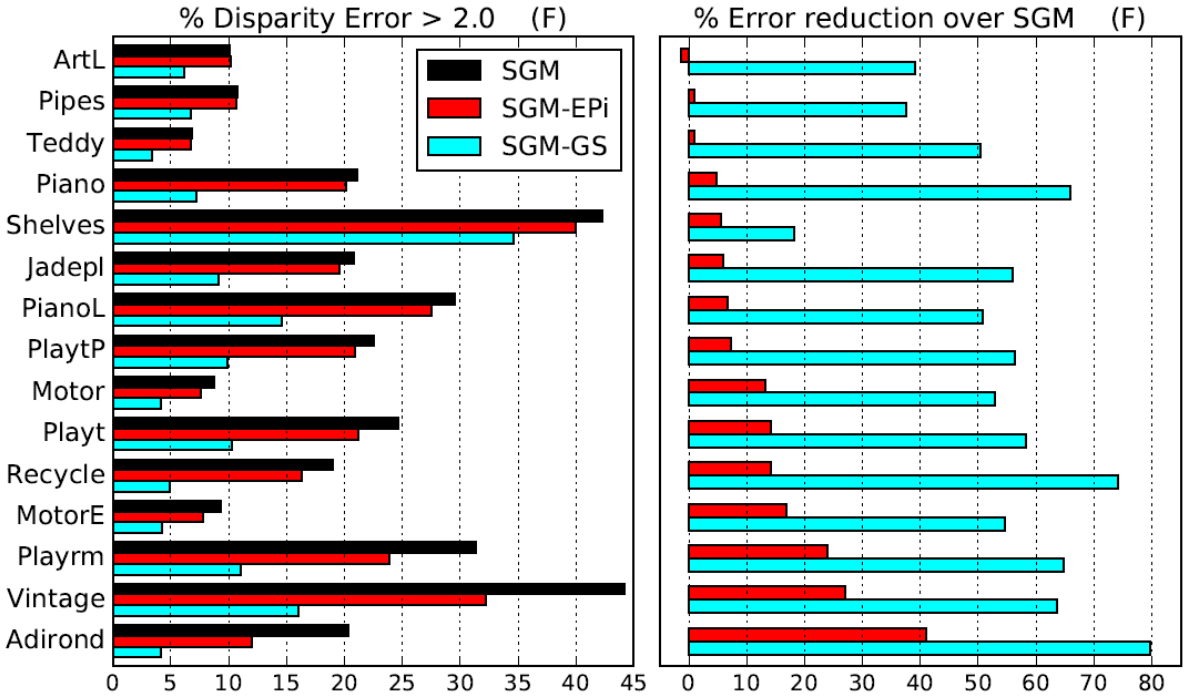
# SGM-P: Where do we get priors?

- Matched features + triangulation
- Matched features + plane fitting
- Low-res matching + plane fitting
- Ground truth oracle
- Semantic analysis
- Manhattan-world assumptions

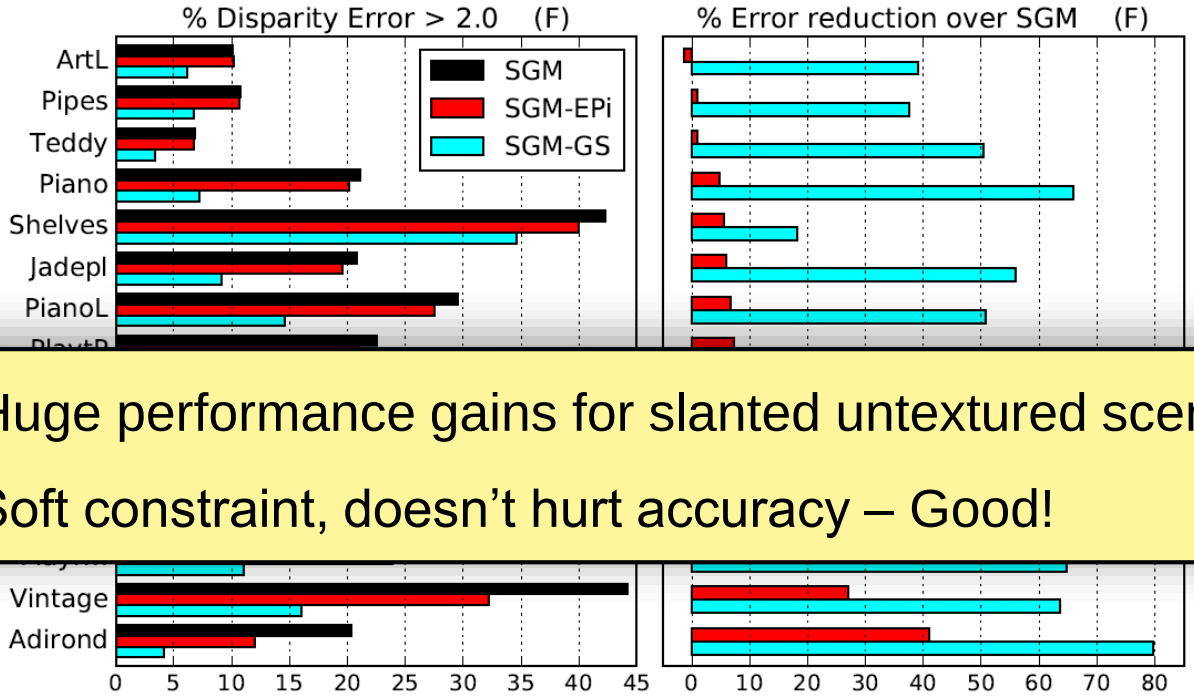
# SGM-P: Results



# SGM-P: Results

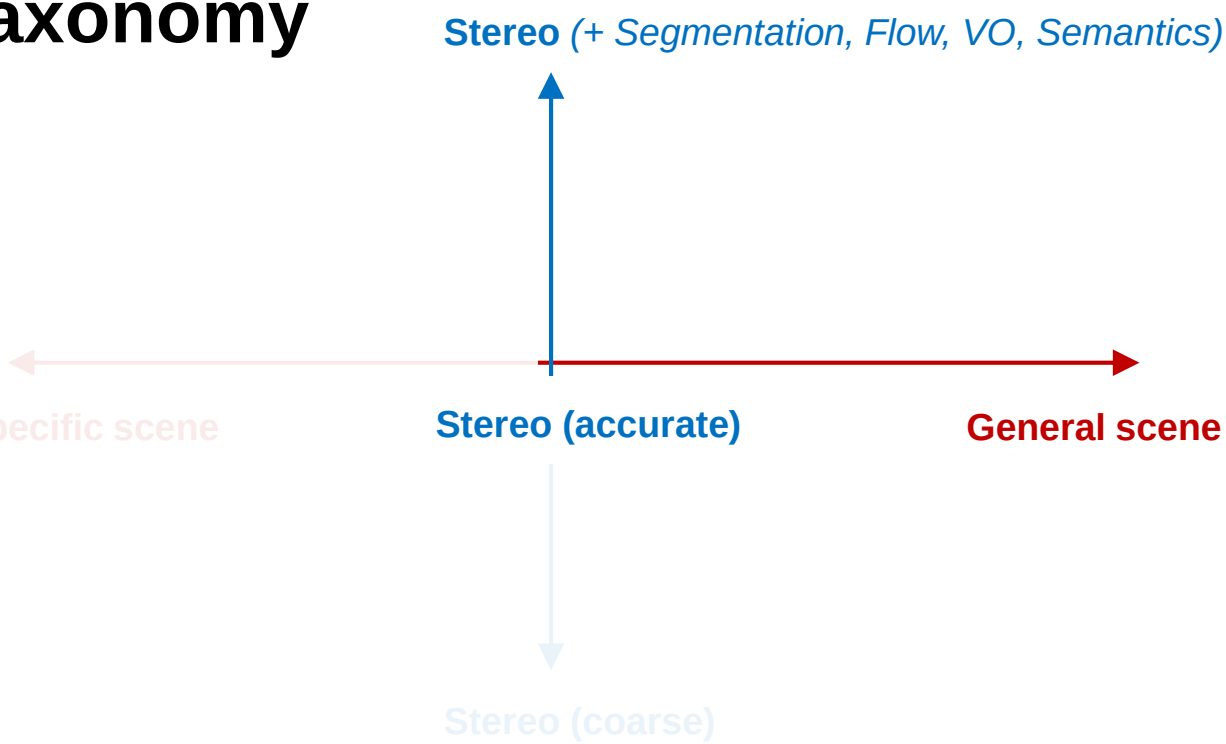


# SGM-P: Results




- Huge performance gains for slanted untextured scenes
- Soft constraint, doesn't hurt accuracy – Good!

# Taxonomy


Stereo (+ Segmentation, Flow, VO, Semantics)
Specific scene
Stereo (accurate)
General scene
Stereo (coarse)

# Joint estimation of …

- Stereo **+** Segmentation (2011)

- Semantic Correspondence **+** Co-Segmentation (2016)

- Scene Flow (2017)

  + Ego-motion

  + Motion Segmentation

# Object Stereo [Bleyer+ 2011]

- Joint recovery of disparity and segmentation in both views
- Model: *Objects in scene, each have*
  - Color model (GMM)
  - Surface model (plane + parallax)
- Assumptions:
  - Color distribution is *compact*
  - Object surface *close to* a 3D plane

# Object Stereo [Bleyer+ 2011]



*Minimize:*

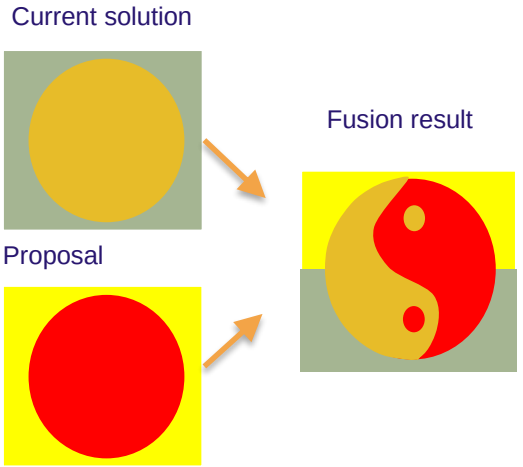
$$E(D,O)$$
$$= E_{photo}(D) + E_{color}(O) + E_{sm-D}(D) + E_{sm-O}(O) + E_{mdl}(D,O) + \ldots$$

# Object Stereo [Bleyer+ 2011]

*Minimize:*

$$E(D,O) = E_{photo}(D) + E_{color}(O) + E_{sm-D}(D) + E_{sm-O}(O) + E_{mdl}(D,O) + \ldots$$

- Proposal generation
- Merge proposals optimally
  - MRF Fusion moves (non-submodular graph cuts)
  - Quadratic Pseudo Boolean Optimization (QPBO)

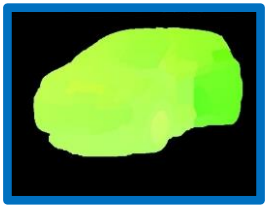


# Joint Correspondence and Co-segmentation

[Taniai, Sinha, Sato 2016]

Source

Target



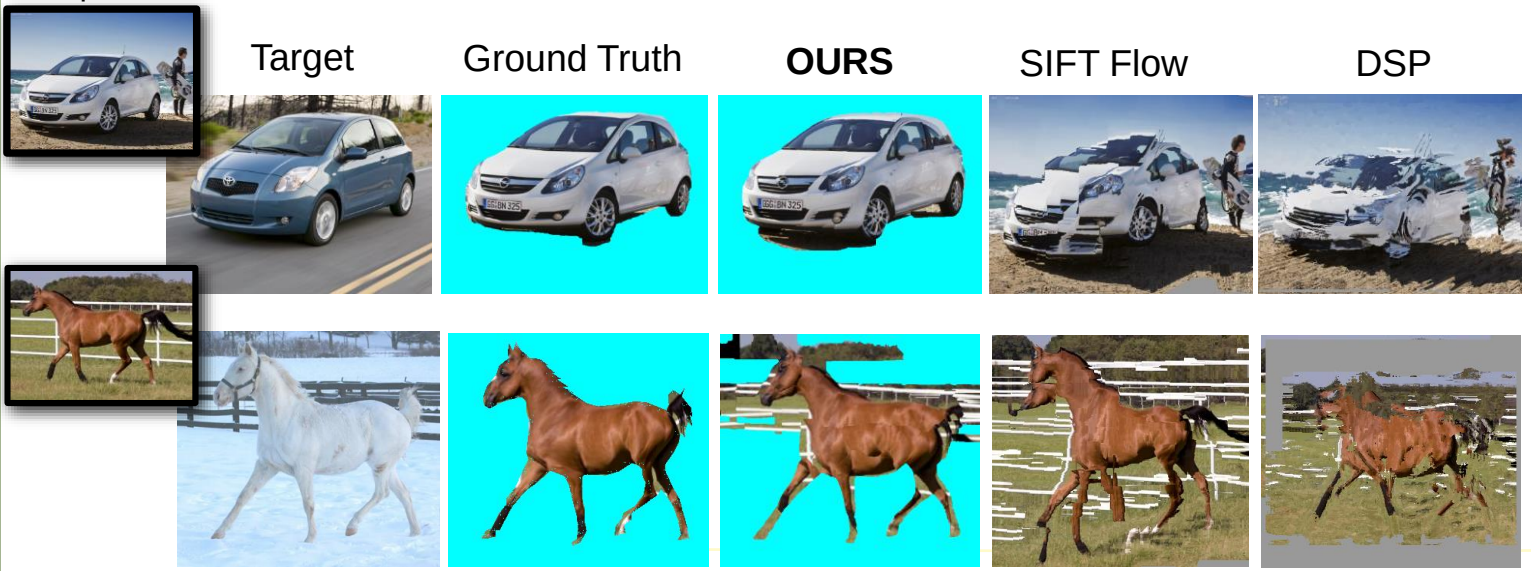
Mask + Flow

Warped Source Image

- **Input:** Images with semantically related objects (different instances)
- **Output:** Common object regions and dense correspondence associated with these regions.

# Joint Correspondence and Co-segmentation

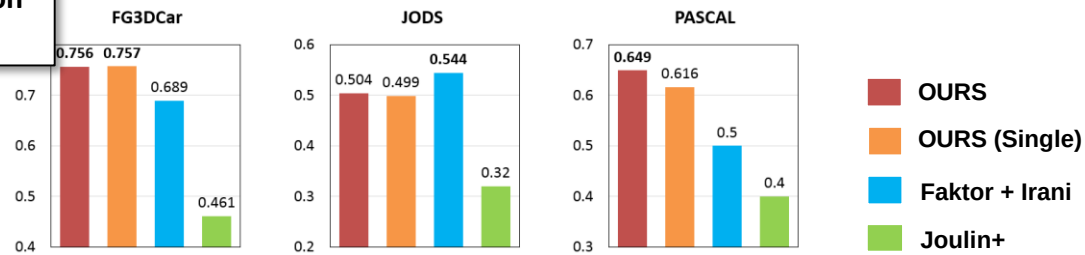
[Taniai, Sinha, Sato 2016]



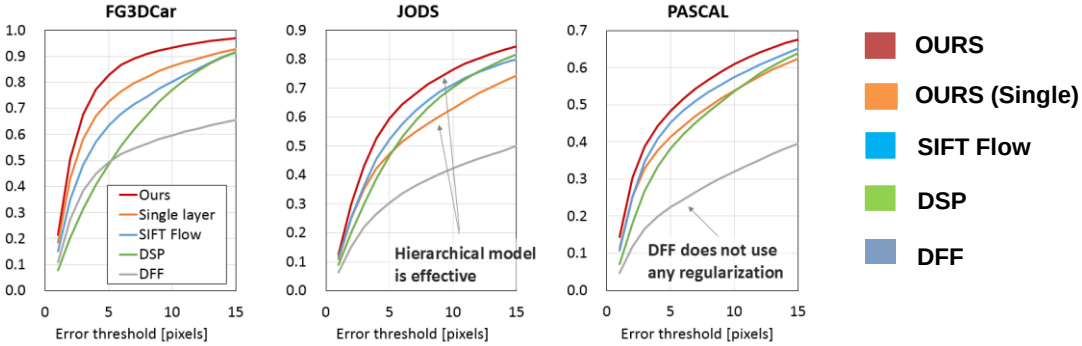
# Joint Correspondence and Co-segmentation

[Taniai, Sinha, Sato 2016]

**Segmentation Accuracy**

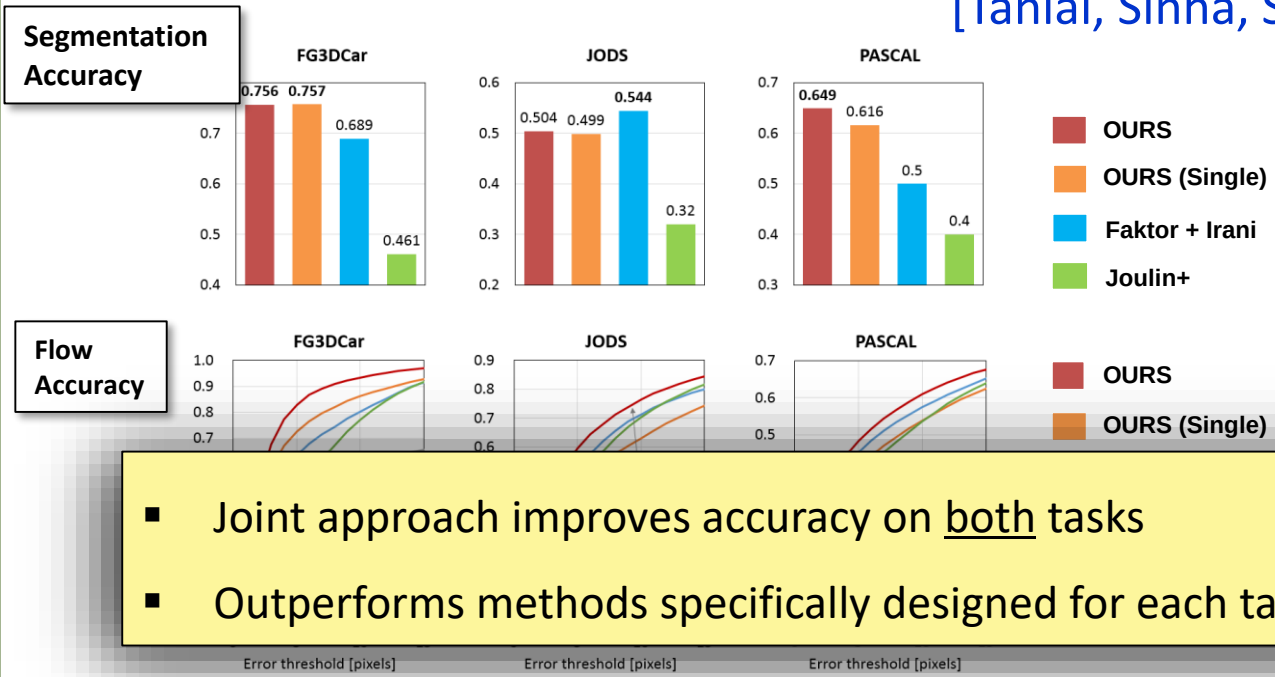



**Flow Accuracy**



# Joint Correspondence and Co-segmentation

[Taniai, Sinha, Sato 2016]




- Joint approach improves accuracy on <u>both</u> tasks
- Outperforms methods specifically designed for each task

## Fast Multi-frame Stereo Scene Flow with Motion Segmentation

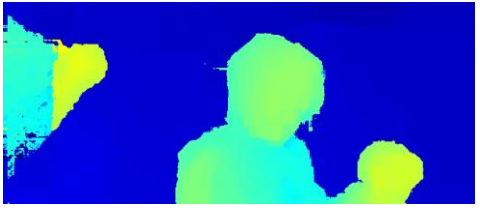
Taniai, Sinha, Sato 2017

**Input:**
**Stereo Video**

Left

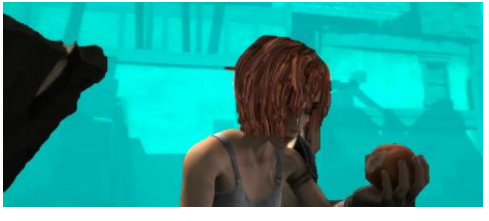
Right

**Output**



Disparity Map

Optical Flow



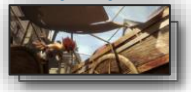
Moving object segmentation

# Fast Multi-frame Stereo Scene Flow with Motion Segmentation

Taniai, Sinha, Sato 2017

**Input**

$I_t^0, I_t^1$



**Binocular stereo**

$\widetilde{\mathcal{D}}$

Init. disparity

$I_t^0, I_{t+1}^0$ + $\widetilde{\mathcal{D}}$

**Visual odometry**

**P**

**Ego-motion**

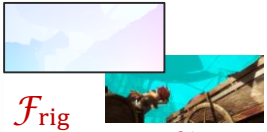
**Epipolar stereo**

$\mathcal{D}$

**Disparity**

$I_{t\pm1}^{0,1}, I_t^0, I_t^1$ + **P**

**Initial motion segmentation**



$\mathcal{F}_{\text{rig}}$

Rigid flow

$\widetilde{\mathcal{S}}$

Initial segmentation

$I_{t\pm1}^{0,1}, I_t^0, I_t^1$ + **P**, $\mathcal{D}$

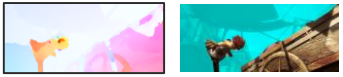
**Optical flow**

$\mathcal{F}_{\text{non}}$

Non-rigid flow

$I_t^0, I_{t+1}^0$ + $\widetilde{\mathcal{S}}$

**Flow fusion**



**Flow**

**Final Segmentation**

$I_t^0, I_{t+1}^0$ + $\mathcal{F}_{\text{rig}}, \mathcal{F}_{\text{non}}$

# Fast Multi-frame Stereo Scene Flow with Motion Segmentation

Taniai, Sinha, Sato 2017

## KITTI 2015 Scene Flow Benchmark (Nov 2016)

| Rank | Method | D1-bg | D1-fg | D1-all | D2-bg | D2-fg | D2-all | Fl-bg | Fl-fg | Fl-all | SF-bg | SF-fg | SF-all | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | PRSM [43] | **3.02** | **10.52** | **4.27** | **5.13** | **15.11** | **6.79** | **5.33** | **17.02** | **7.28** | **6.61** | **23.60** | **9.44** | 300 s |
| 2 | OSF [30] | 4.54 | 12.03 | 5.79 | 5.45 | 19.41 | 7.77 | 5.62 | 22.17 | 8.37 | 7.01 | 28.76 | 10.63 | 50 min |
| 3 | **FSF+MS** (ours) | 5.72 | 11.84 | 6.74 | 7.57 | 21.28 | 9.85 | 8.48 | 29.62 | 12.00 | 11.17 | 37.40 | 15.54 | 2.7 s |
| 4 | CSF [28] | 4.57 | 13.04 | 5.98 | 7.92 | 20.76 | 10.06 | 10.40 | 30.33 | 13.71 | 12.21 | 36.97 | 16.33 | 80 s |
| 5 | PR-Sceneflow [42] | 4.74 | 13.74 | 6.24 | 11.14 | 20.47 | 12.69 | 11.73 | 27.73 | 14.39 | 13.49 | 33.72 | 16.85 | 150 s |
| 8 | PCOF + ACTF [10] | 6.31 | 19.24 | 8.46 | 19.15 | 36.27 | 22.00 | 14.89 | 62.42 | 22.80 | 25.77 | 69.35 | 33.02 | 0.08 s (GPU) |
| 12 | GCSF [8] | 11.64 | 27.11 | 14.21 | 32.94 | 35.77 | 33.41 | 47.38 | 45.08 | 47.00 | 52.92 | 59.11 | 53.95 | 2.4 s |

200 road scenes with multiple moving objects

| Rank | SF-all | Time |
|---|---|---|
| 1 | **9.44** | 300 s |
| 2 | 10.63 | 50 min |
| 3 | 15.54 | 2.7 s |
| 4 | 16.33 | 80 s |
| 5 | 16.85 | 150 s |
| 8 | 33.02 | 0.08 s (GPU) |
| 12 | 53.95 | 2.4 s |

# Summary

- Piecewise planar stereo
  - Coarse, simplified depth maps
  - Good for Image-based rendering
  - Planes (surfaces) as primitives
  - How about higher-level primitives?
    - 3D CAD models
    - 3D shape templates

# Summary

- Extending SGM
  - Reducing the search space (local plane sweeps)
  - Soft orientation prior
- Merit of joint formulations for complementary tasks
  - Stereo + segmentation
  - Semantic correspondence
  - Scene fow from stereoscopic video

# Collaborators

Rick Szeliski
Facebook

Daniel Scharstein,
Middlebury College

Tatsunori Taniai
RIKEN, Tokyo

Adarsh Kowdle
PerceptiveIO

Michael Bleyer
Microsoft

Michael Goesele
TU Darmstadt

Johannes Kopf
Facebook

Drew Steedly
Microsoft

Yoichi Sato
Univ. of Tokyo

Carsten Rother
TU Dresden

Pushmeet Kohli
DeepMind