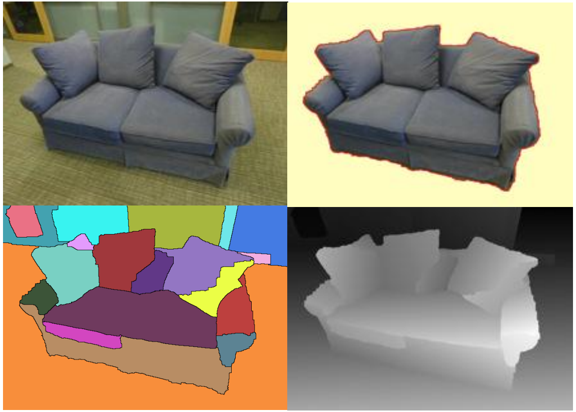
Multiple View Object Cosegmentation using Appearance and Stereo Cues

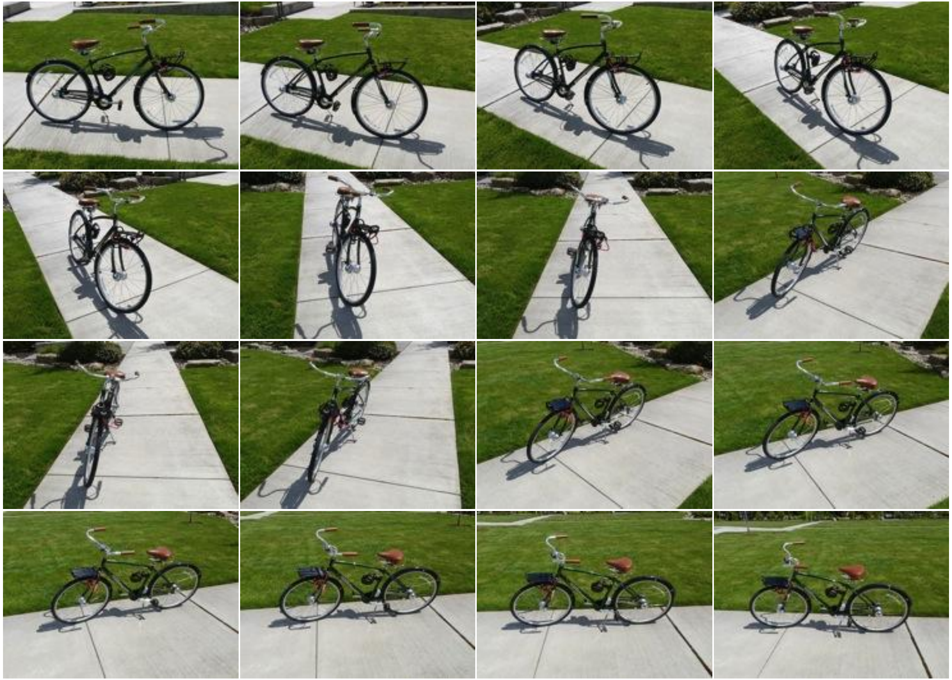
Adarsh Kowdle[1]

Sudipta Sinha[2]

Richard Szeliski[2]

[1]Cornell University, [2]Microsoft Research



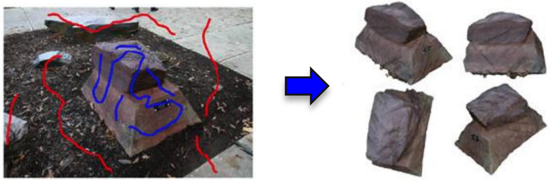


Final result using our approach

Previous Work

- Interactive (co)-segmentation

GrabCut - Rother et. al.
SIGGRAPH ‘04



iCoseg - Batra et. al.
IJCV ‘11

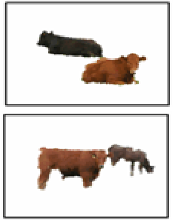
iModel - Kowdle et. al.
ECCV - RMLE ‘10

- Unsupervised cosegmentation

Mukherjee et. al.
CVPR ‘11

Vicente et. al.
CVPR ‘11

Kim et. al.
ICCV ‘11, CVPR ‘12



Joulin et. al.
CVPR ‘12

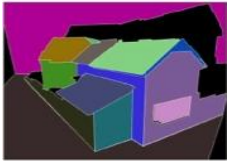
Previous Work

- Unsupervised 3D reconstruction *and* cosegmentation

Furukawa et. al. (PMVS)
CVPR ‘07

Campbell et. al.
BMVC ‘07, CVMP ‘11

- Piecewise planar stereo and low-level segmentation



Birchfield et. al.
ICCV ‘99

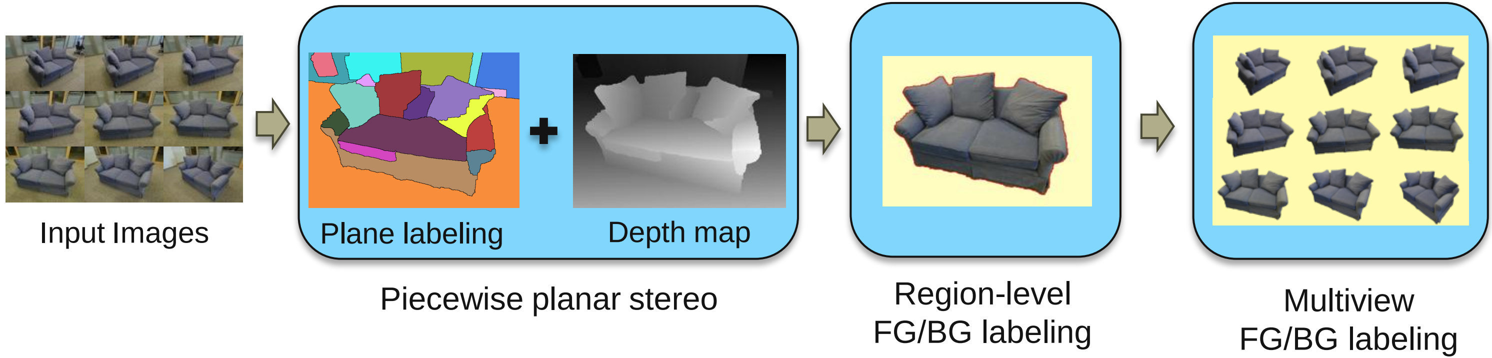
Sinha et. al.
ICCV ‘09

Bleyer et. al.
CVPR ‘11

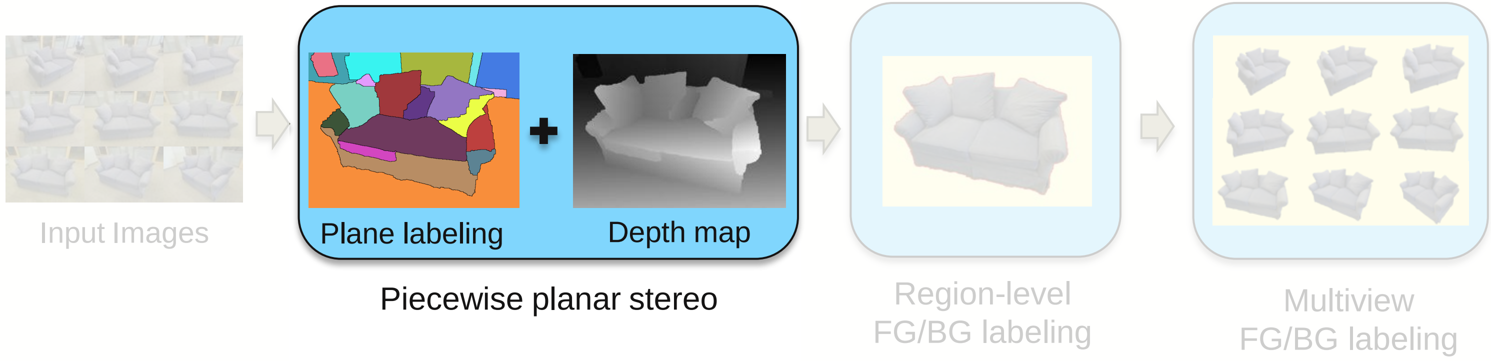
Contributions

- Unsupervised object cosegmentation algorithm
  - exploits stereo and appearance cues
- Extend prior work on piecewise planar stereo
  - robust to scenarios where stereo matching is unreliable

Overview

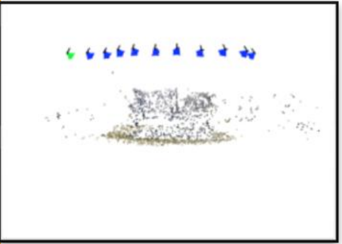

Input Images
Plane labeling
Depth map
Piecewise planar stereo
Region-level
FG/BG labeling
Multiview
FG/BG labeling

Overview

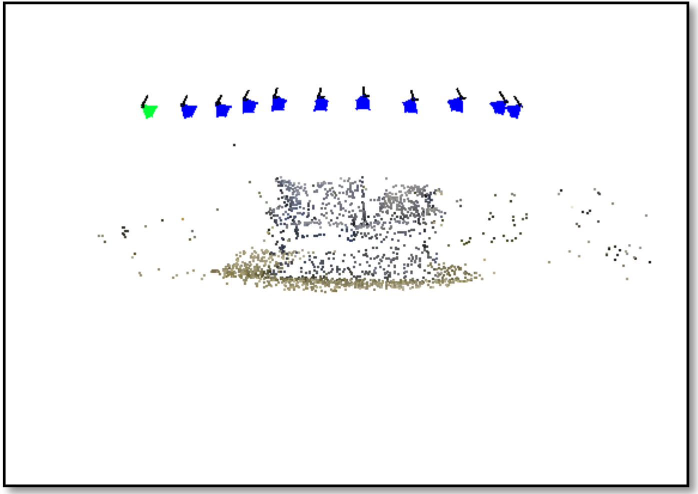

Input Images
Plane labeling
Depth map
Piecewise planar stereo
Region-level
FG/BG labeling
Multiview
FG/BG labeling

Piecewise planar stereo

Input Images



SfM

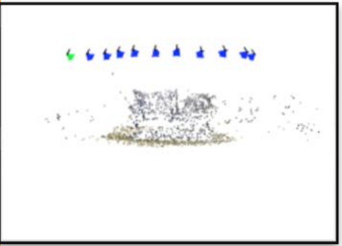


Structure from Motion (SfM)

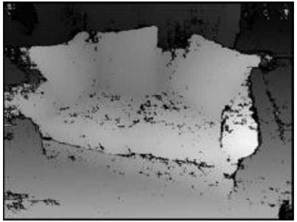
Piecewise planar stereo

Input Images

SGM Stereo



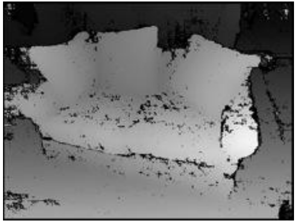
SfM



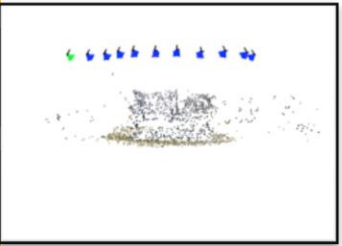
Semi-global matching
(SGM Stereo)
[Hirschmüller 2008]

Piecewise planar stereo

Input Images



SGM Stereo

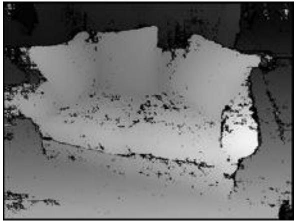


SfM

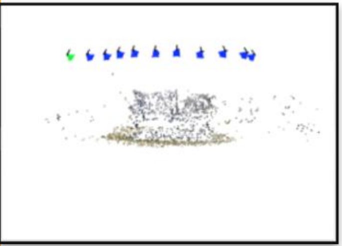
Plane hypotheses

Piecewise planar stereo

Input Images

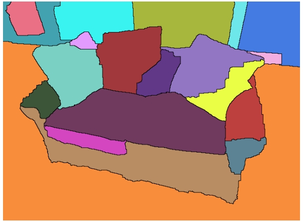


SGM Stereo

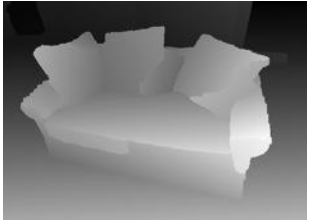


SfM

Plane hypotheses



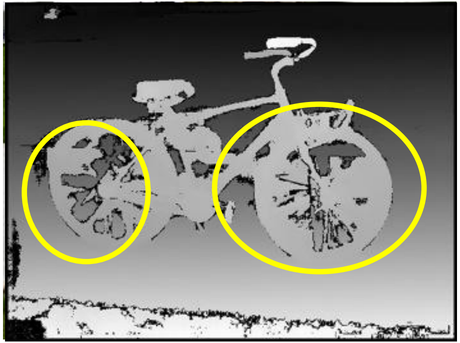
Plane labeling



Resulting depth map

Piecewise planar stereo

Piecewise planar stereo



SGM Stereo
Bicycle sequence

Piecewise planar depthmap
Only stereo cues
Sinha *et. al.* 2009

Our approach
Appearance and stereo cues

Piecewise planar stereo

Pixel level MRF

Grid graph over all pixels p

$$l_p \in \Pi = \{\pi_i\}$$

Each plane π_i is parameterized by

1. 3D plane equation
2. Appearance model ($\mathbf{A}_i$)

Piecewise planar stereo

$$E(L) = \sum_{p \in P} E_p^A(l_p) + \lambda_G \sum_{p \in P} c_p E_p^G(l_p) + \lambda_S \sum_{(p,q) \in \mathcal{N}} E_{pq}\left(l_p, l_q\right)$$

$$l_p \in \Pi = \{\pi_i\}$$

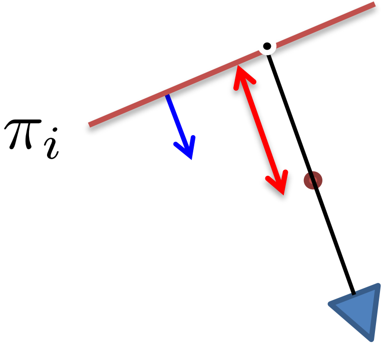
Each plane π_i is parameterized by

1. 3D plane equation
2. Appearance model ($\mathbf{A}_i$)

Piecewise planar stereo

$$E(L) = \sum_{p \in P} E_p^A(l_p) + \lambda_G \sum_{p \in P} c_p \boxed{E_p^G(l_p)} + \lambda_S \sum_{(p,q) \in \mathcal{N}} E_{pq}(l_p, l_q)$$

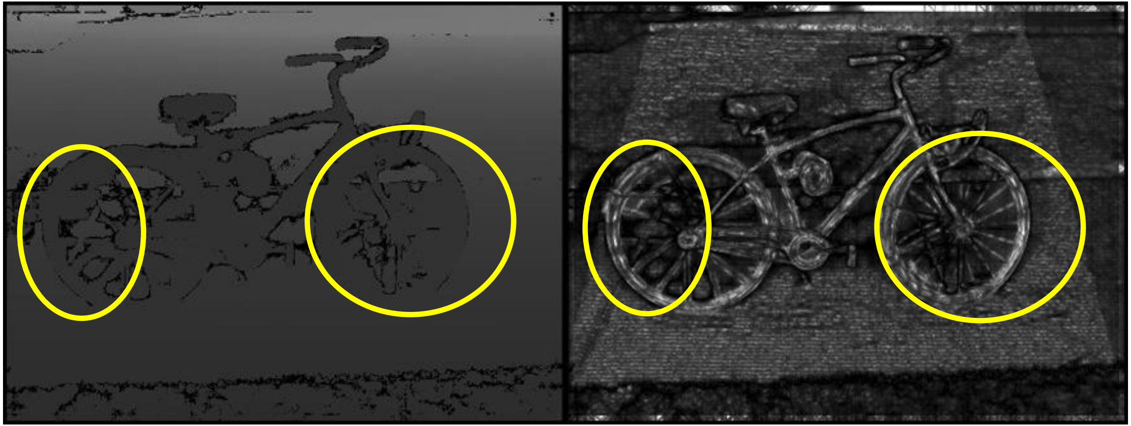
Geometric unary term



Piecewise planar stereo

$$E(L) = \sum_{p \in P} E_p^A(l_p) + \lambda_G \sum_{p \in P} \boxed{c_p} E_p^G(l_p) + \lambda_S \sum_{(p,q) \in \mathcal{N}} E_{pq}(l_p, l_q)$$

Per-pixel confidence



SGM Stereo Confidence map

Piecewise planar stereo

$$E(L) = \sum_{p \in P} \boxed{E_p^A(l_p)} + \lambda_G \sum_{p \in P} c_p E_p^G(l_p) + \lambda_S \sum_{(p,q) \in \mathcal{N}} E_{pq}(l_p, l_q)$$

Appearance unary term

Appearance model
Lab features (GMM)
$\mathbf{A}_i$

$$E_p^A(l_p = \pi_i) = -log(p(\mathbf{x}|\mathbf{A}_i))$$

Per-region color models vs.
global color models

Piecewise planar stereo

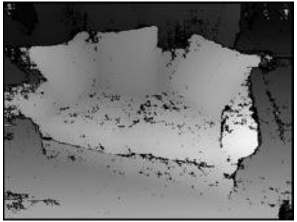
$$E(L) = \sum_{p \in P} E_p^A(l_p) + \lambda_G \sum_{p \in P} c_p E_p^G(l_p) + \lambda_S \sum_{(p,q) \in \mathcal{N}} \boxed{E_{pq}\left(l_p, l_q\right)}$$

Pairwise term

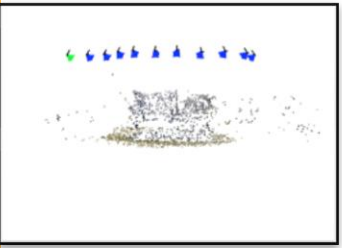
Contrast sensitive Potts Model

Piecewise planar stereo

Input Images



Three-view stereo

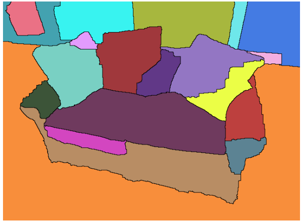


SfM

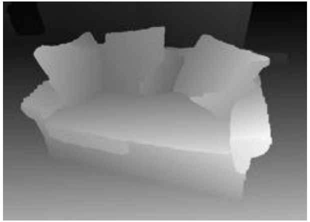
Plane hypotheses

Iterative graph cut with alpha-expansion

(Typically 2-3 iterations)



Plane labels

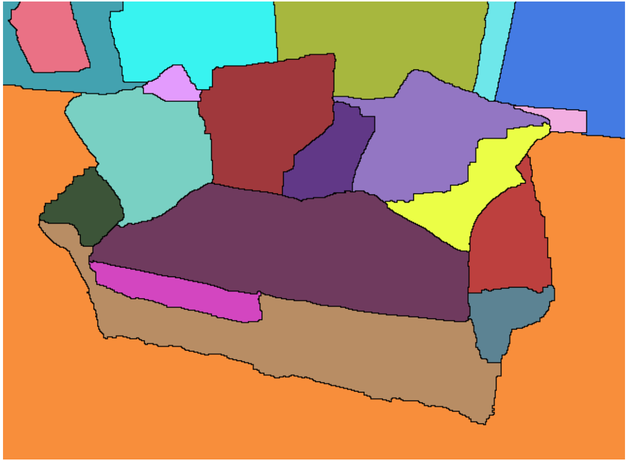


Piecewise planar depth map

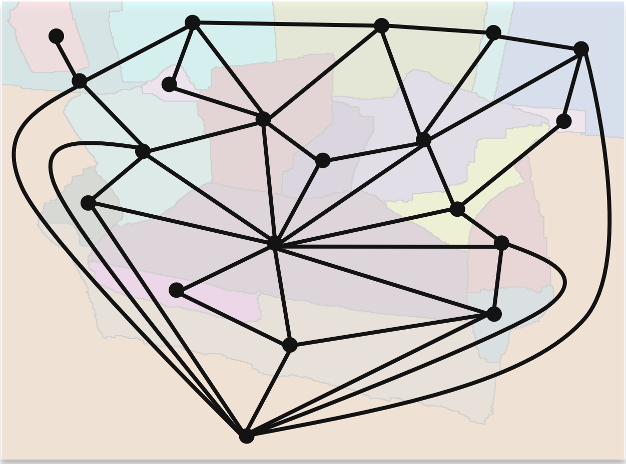
Overview

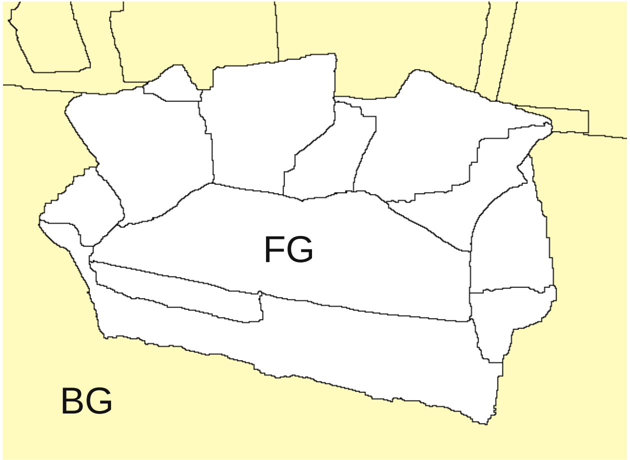

Input Images
Plane labeling
Depth map
Piecewise planar stereo
Region-level
FG/BG labeling
Multiview
FG/BG labeling

Region-level FG/BG labeling



Region-level FG/BG labeling





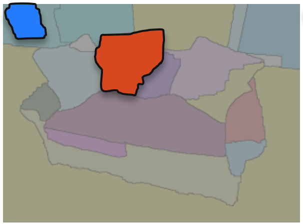
Region level labeling

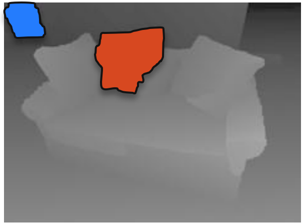
Region-level FG/BG labeling

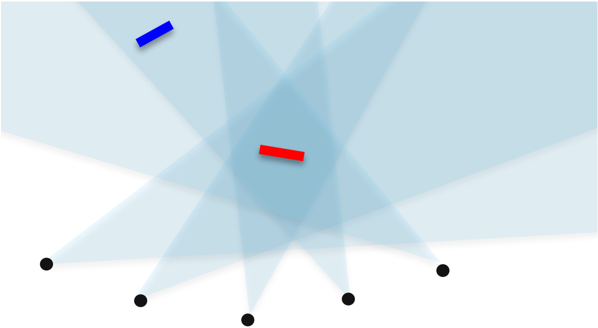
Appearance → $$E(F_r^a) = \sum_{r \in R} \boxed{E_r^O(f_r)} + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^a(f_r, f_t)$$

Geometry → $$E(F_r^g) = \sum_{r \in R} \boxed{E_r^O(f_r)} + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^g(f_r, f_t)$$

Objectness term





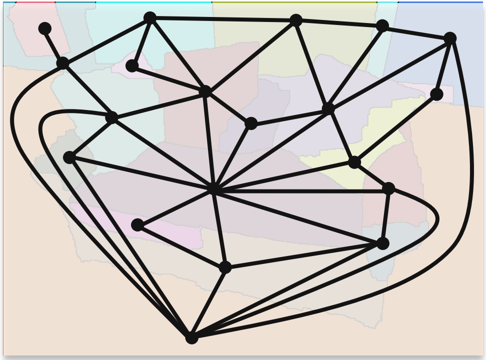


Region-level FG/BG labeling

Appearance → $$E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} \boxed{E_{rt}^a(f_r, f_t)}$$

Geometry → $$E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^g(f_r, f_t)$$

Appearance-based compatibility

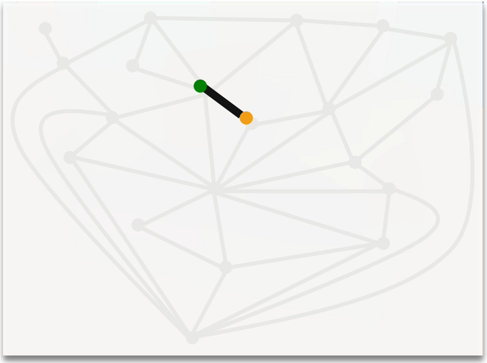


Region-level FG/BG labeling

Appearance → $$E(F_r^a) = \sum_{r\in R} E_r^O(f_r) + \sum_{(r,t)\in \mathcal{N}_R} \boxed{E_{rt}^a(f_r, f_t)}$$

Geometry → $$E(F_r^g) = \sum_{r\in R} E_r^O(f_r) + \sum_{(r,t)\in \mathcal{N}_R} E_{rt}^g(f_r, f_t)$$

Appearance-based compatibility

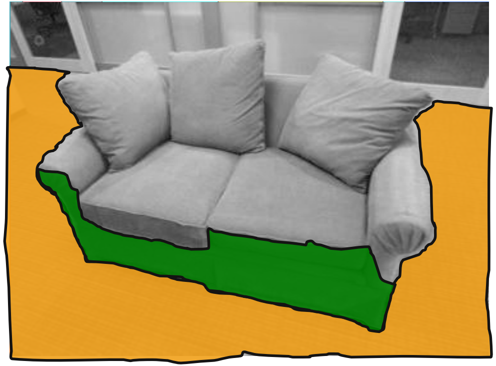


Region-level FG/BG labeling

Appearance → $E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} \boxed{E_{rt}^a(f_r, f_t)}$

Geometry → $E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^g(f_r, f_t)$

Appearance-based compatibility

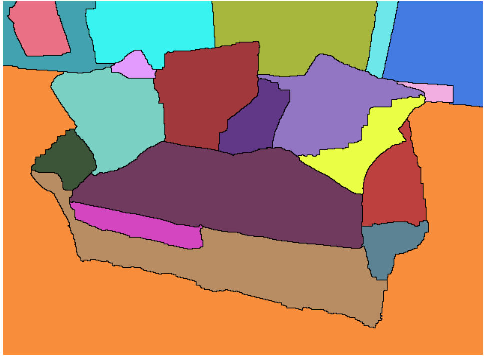


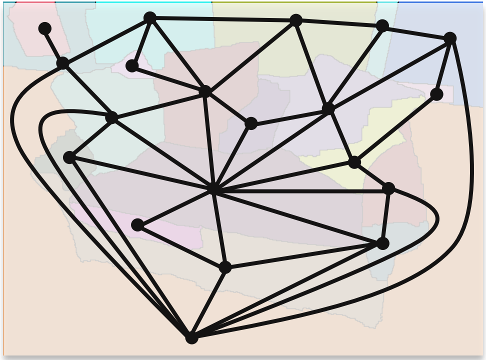
Region-level FG/BG labeling

Appearance → $E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^a(f_r, f_t)$

Geometry → $E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} \boxed{E_{rt}^g(f_r, f_t)}$

Depth-based compatibility



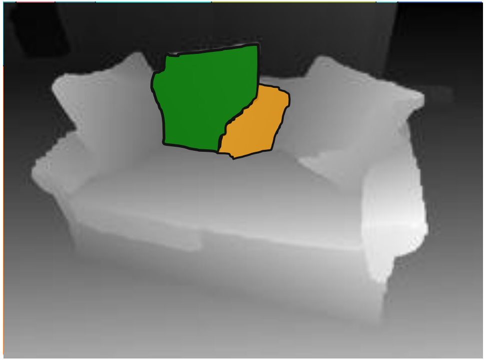


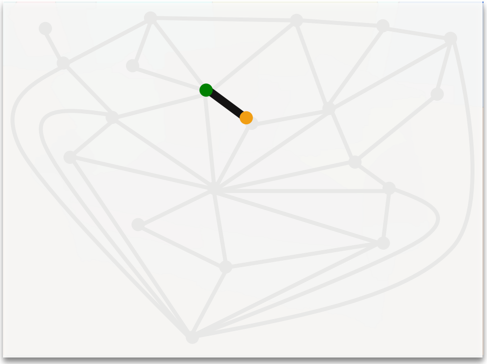
Region-level FG/BG labeling

Appearance → $E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^a(f_r, f_t)$

Geometry → $E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} \boxed{E_{rt}^g(f_r, f_t)}$

Depth-based compatibility



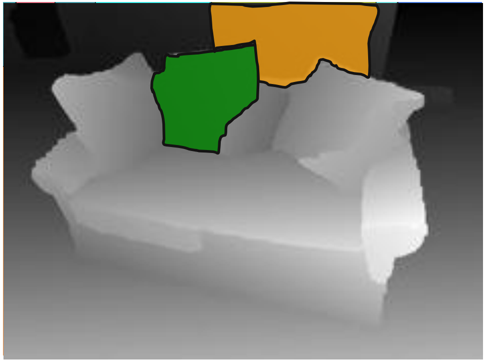


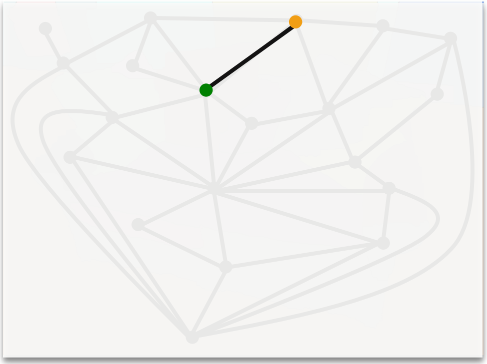
Region-level FG/BG labeling

Appearance → $E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^a(f_r, f_t)$

Geometry → $E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} \boxed{E_{rt}^g(f_r, f_t)}$

Depth-based compatibility





Region-level FG/BG labeling

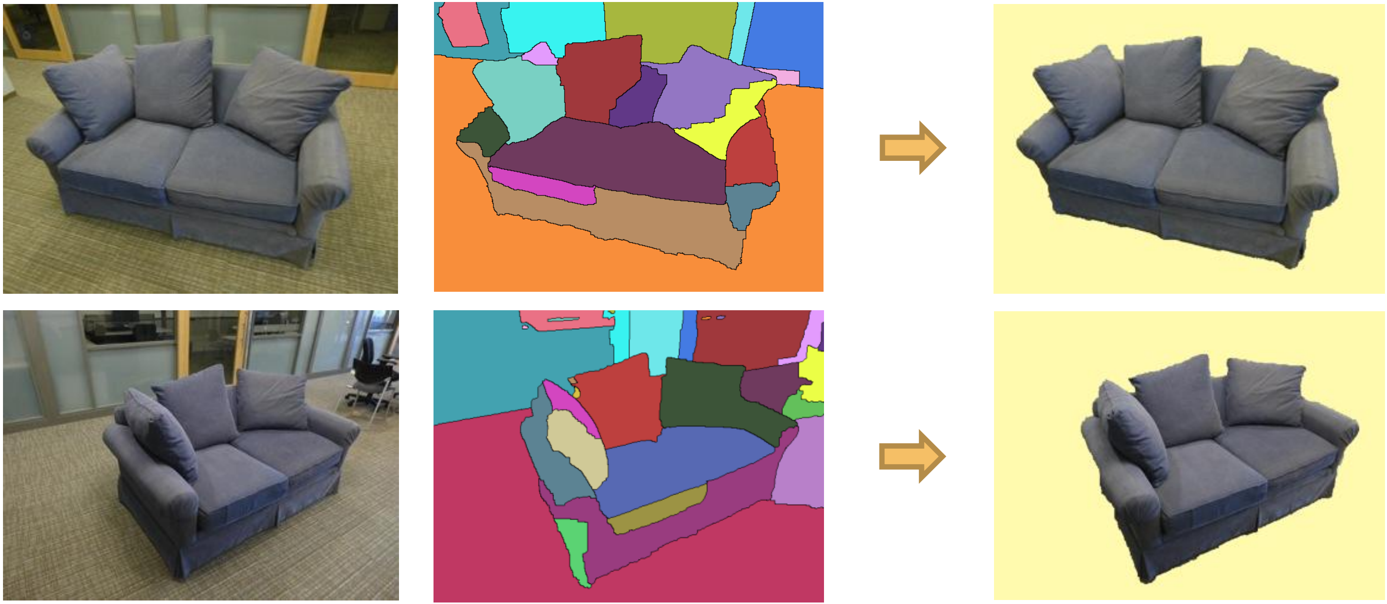
Appearance → $$E(F_r^a) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^a(f_r, f_t)$$

Geometry → $$E(F_r^g) = \sum_{r \in R} E_r^O(f_r) + \sum_{(r,t) \in \mathcal{N}_R} E_{rt}^g(f_r, f_t)$$

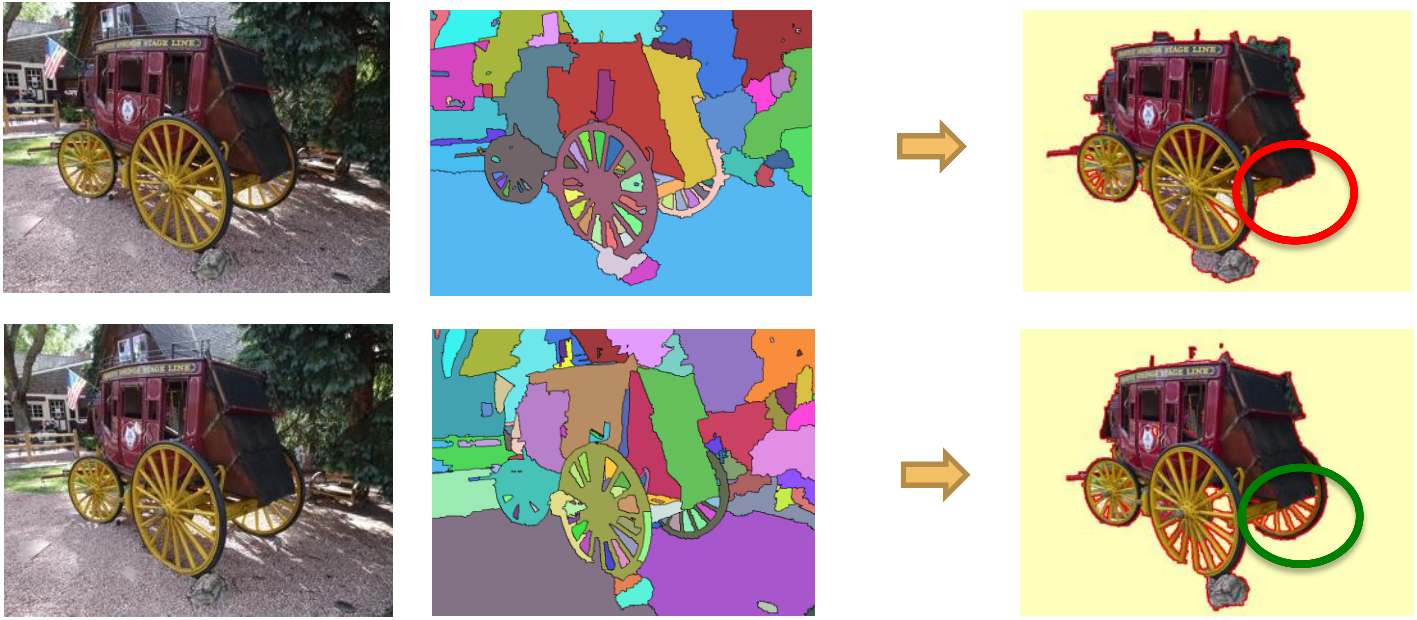
Graph cut on each energy function independently
to obtain MAP labels

Region labeled FG if either solutions label region FG

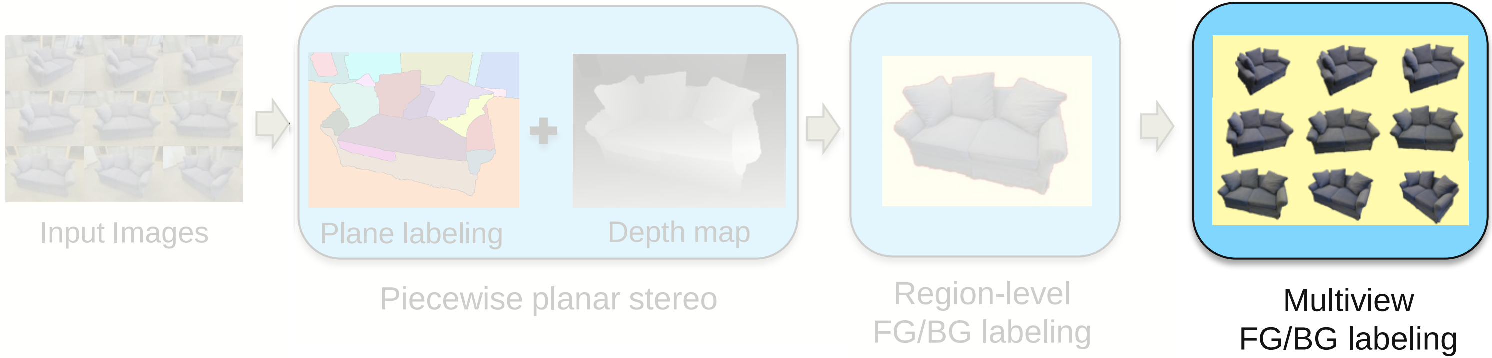
Region-level FG/BG labeling



Region-level FG/BG labeling



Overview


Input Images
Plane labeling
Depth map
Piecewise planar stereo
Region-level
FG/BG labeling
Multiview
FG/BG labeling

Multiview FG/BG labeling

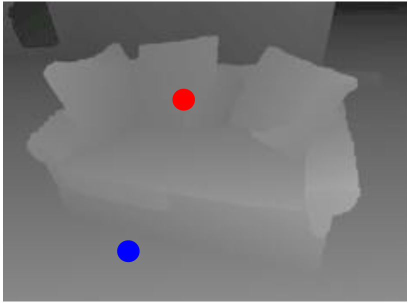
Pixel level MRF

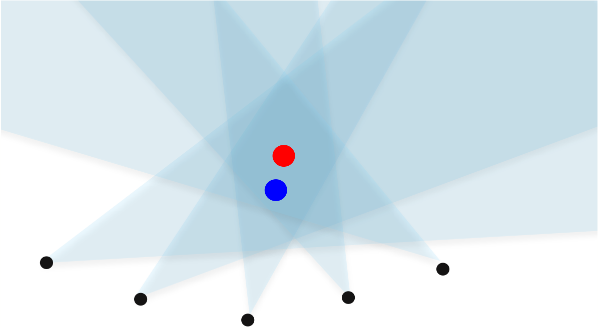
Grid graph over all pixels p

Multiview FG/BG labeling

$$E(F) = \sum_{p \in P} \boxed{E_p^O(f_p)} + \sum_{p \in P} E_p^A(f_p) + \sum_{(p,q) \in \mathcal{N}} E_{pq}(f_p, f_q)$$

Objectness term

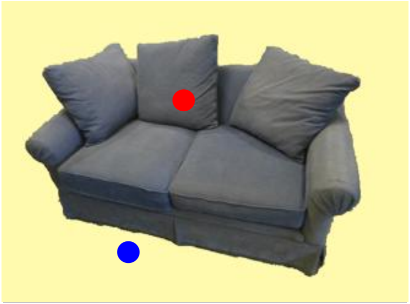




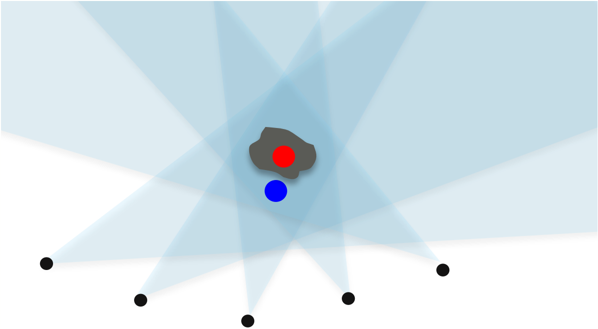
Multiview FG/BG labeling

$$E(F) = \sum_{p \in P} \boxed{E_p^O(f_p)} + \sum_{p \in P} E_p^A(f_p) + \sum_{(p,q) \in \mathcal{N}} E_{pq}(f_p, f_q)$$

Objectness term



Region-level FG/BG labeling



Multiview FG/BG labeling

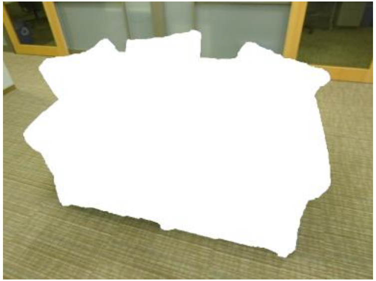
$$E(F) = \sum_{p \in P} E_p^O(f_p) + \sum_{p \in P} \boxed{E_p^A(f_p)} + \sum_{(p,q) \in \mathcal{N}} E_{pq}(f_p, f_q)$$

Appearance unary term

$$\mathbf{A} = \{\mathbf{A}_f, \mathbf{A}_b\}$$

Region-level FG/BG labeling

FG



BG

Multiview FG/BG labeling

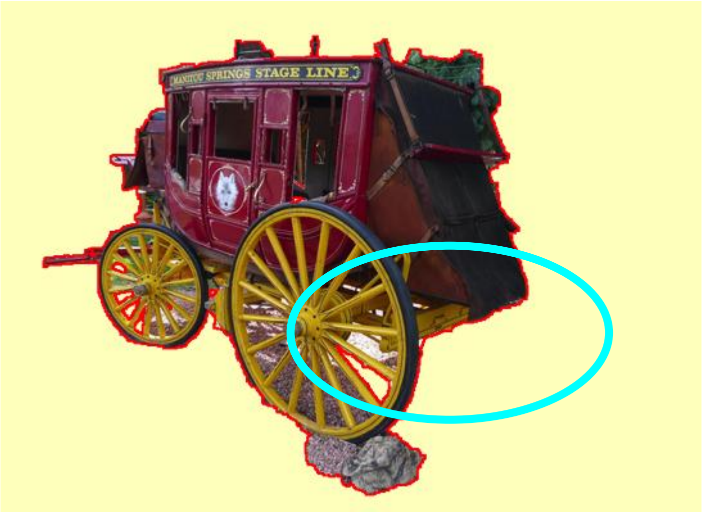
$$E(F) = \sum_{p \in P} E_p^O(f_p) + \sum_{p \in P} E_p^A(f_p) + \sum_{(p,q) \in \mathcal{N}} \boxed{E_{pq}(f_p, f_q)}$$

Pairwise term

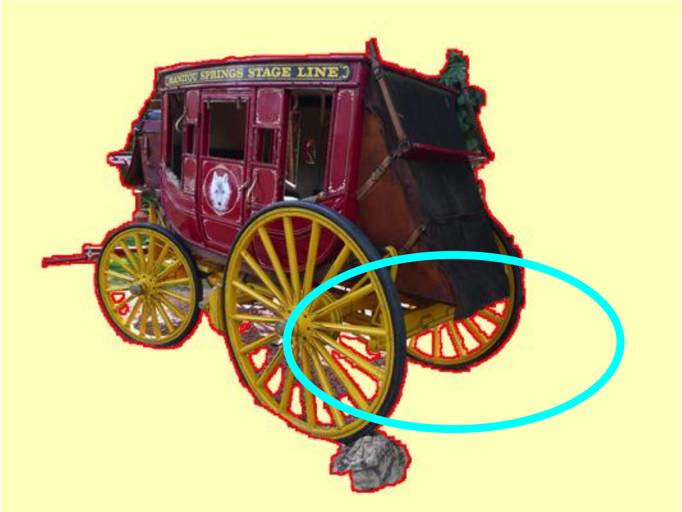
Contrast sensitive Potts Model

Multiview FG/BG labeling

Graph cut to obtain MAP labels

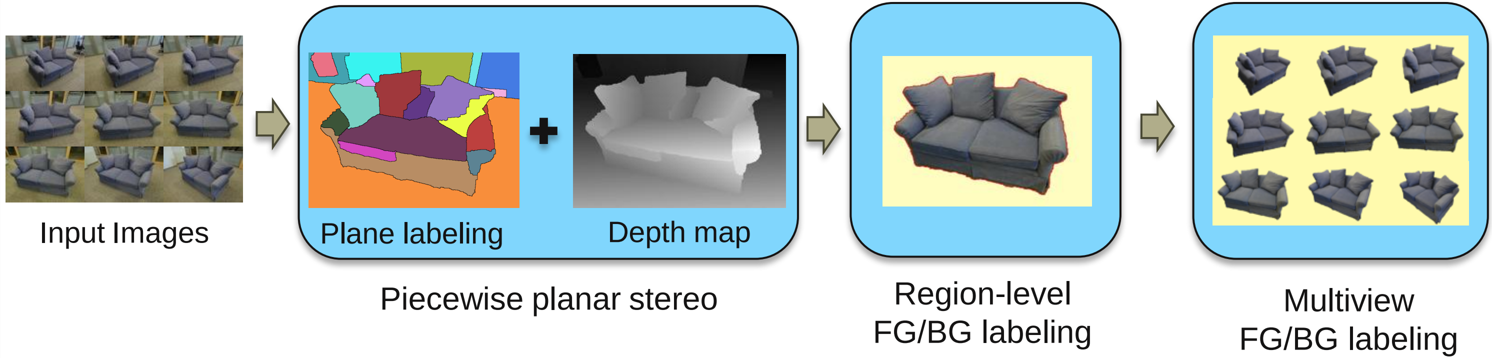


Region-level FG/BG labeling

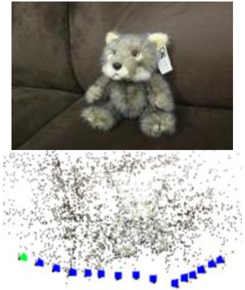
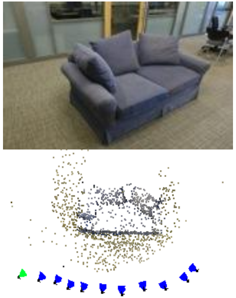


Multiview FG/BG labeling

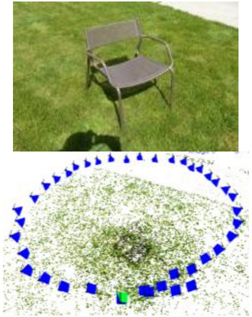
Overview


Input Images
Plane labeling
Depth map
Piecewise planar stereo
Region-level
FG/BG labeling
Multiview
FG/BG labeling

Datasets



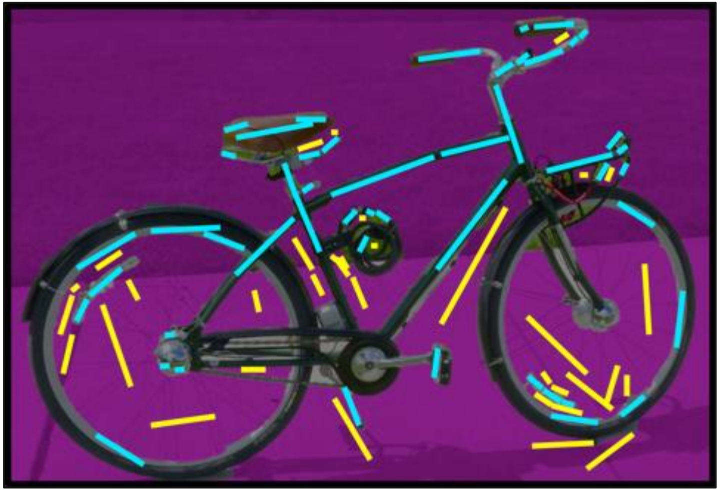



Quantitative results

Ground truth: Manually labeled using GrabCut

Quantitative results

Ground truth: Manually labeled using GrabCut



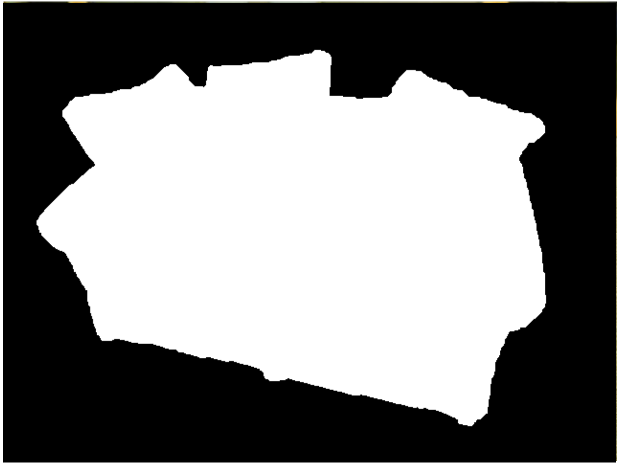
4 minutes

Quantitative results

Evaluation metric

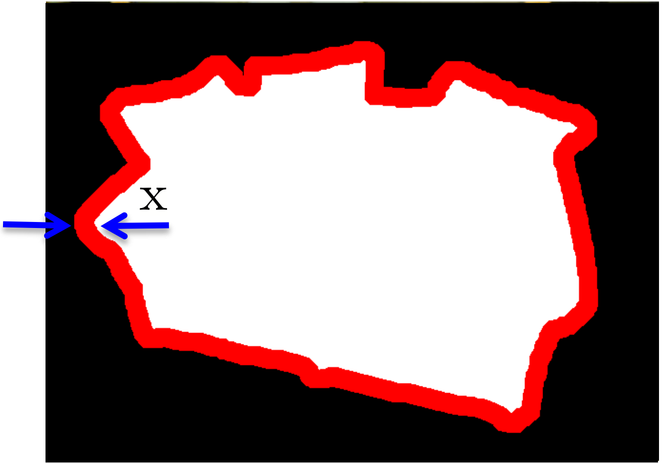
Quantitative results

Evaluation metric



Quantitative results

Evaluation metric



$$\text{Acc-x} = \frac{\text{Intersection}}{\text{Union}}$$

Comparisons

GroundTruth
GrabCut, Rother et. al.
[SIGGRAPH '04]

Vicente et. al.
[CVPR '11]

PMVS
Furukawa et. al.
[CVPR '10]

Ours

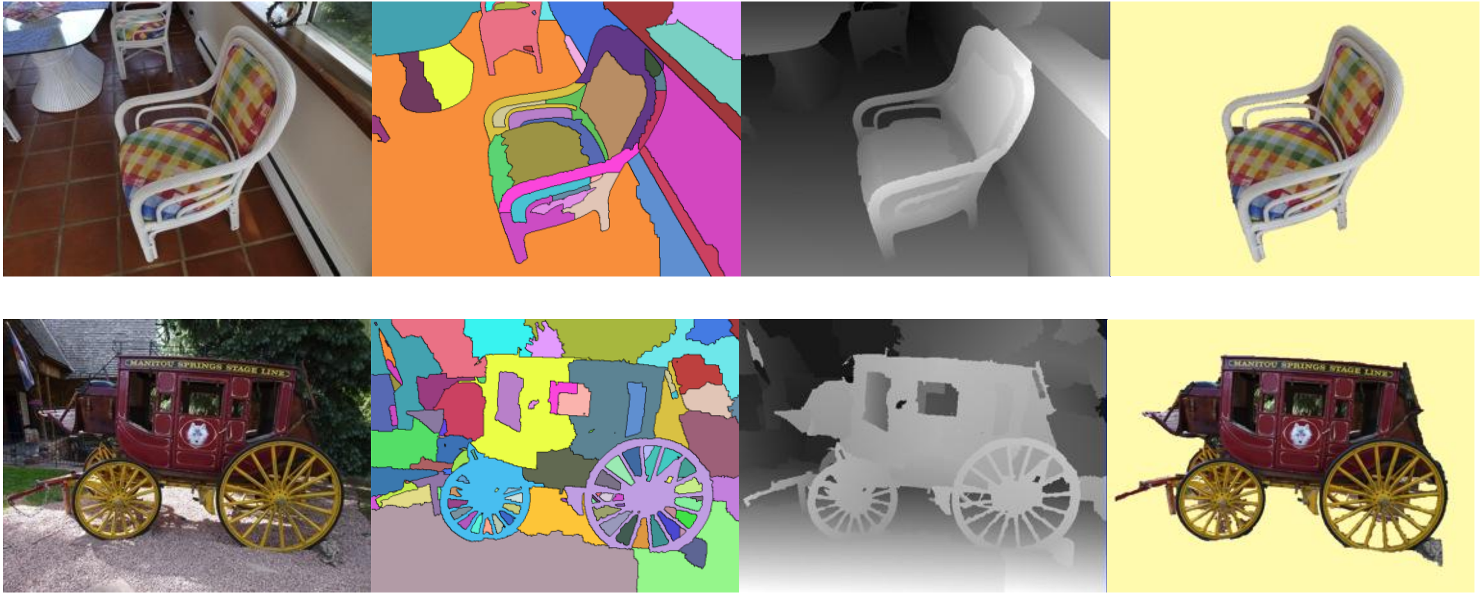
Acc-10	56.5 ± 3.1	56.2 ± 1.8	$\mathbf{89.8 \pm 3.8}$
Acc-Full	81.7 ± 8.9	89.1 ± 3.9	$\mathbf{98.0 \pm 0.5}$

Comparisons

Name		Vicente'11	PMVS'12	Ours
BIKE	Acc-10	68.1 ± 6.7	61.0 ± 3.9	$\mathbf{90.0 \pm 4.9}$
	Acc-Full	88.9 ± 6.3	96.0 ± 1.8	$\mathbf{99.1 \pm 0.7}$
BICYCLE	Acc-10	56.5 ± 3.1	56.2 ± 1.8	$\mathbf{89.8 \pm 3.8}$
	Acc-Full	81.7 ± 8.9	89.1 ± 3.9	$\mathbf{98.0 \pm 0.5}$
CHAIR1	Acc-10	73.3 ± 4.8	72.7 ± 2.1	$\mathbf{93.9 \pm 3.1}$
	Acc-Full	86.9 ± 7.8	96.6 ± 0.4	$\mathbf{99.2 \pm 0.4}$
CAR	Acc-10	74.4 ± 5.3	59.6 ± 4.3	$\mathbf{83.2 \pm 1.1}$
	Acc-Full	91.8 ± 4.3	91.2 ± 5.5	$\mathbf{97.9 \pm 0.6}$

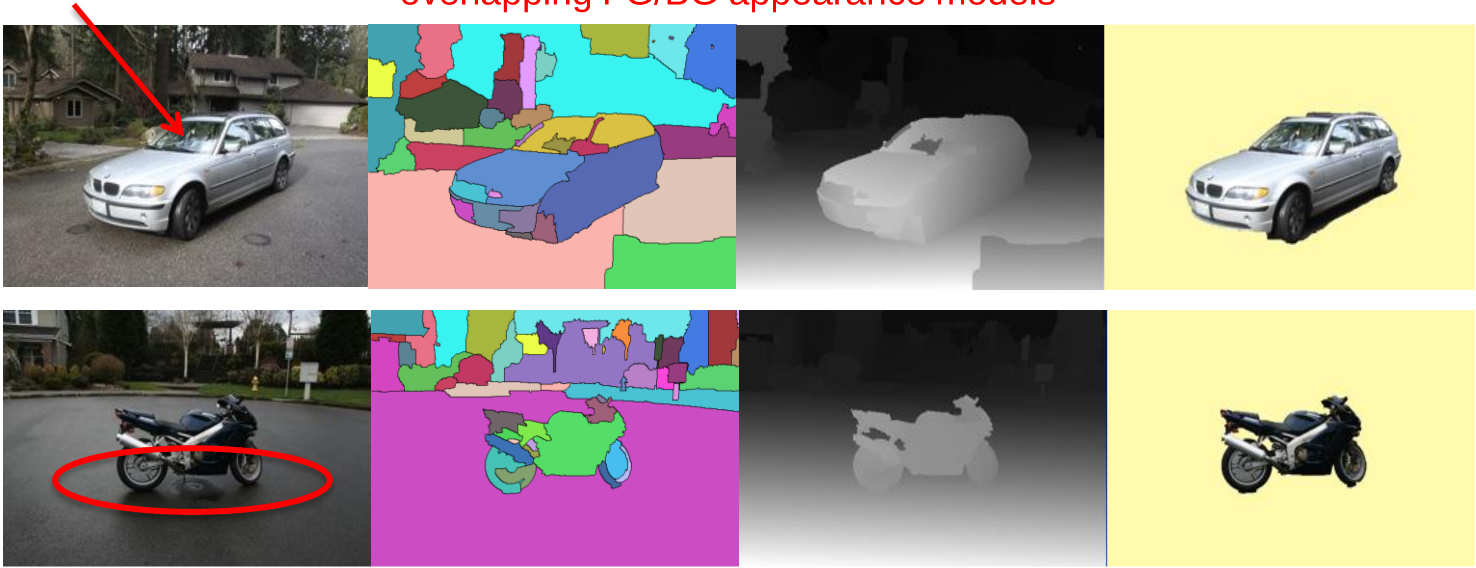
Qualitative results

Complicated object structures modeled via piecewise planar proxies



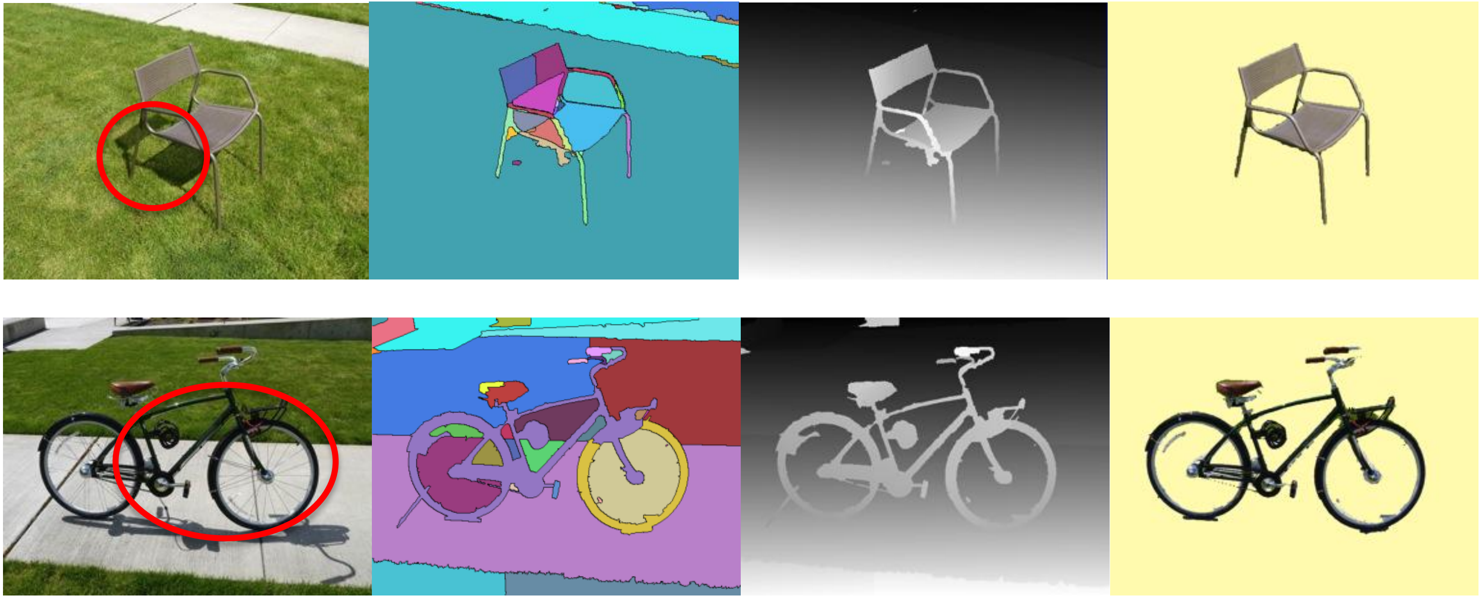
Qualitative results

Irregularities such as specular surfaces and overlapping FG/BG appearance models



Qualitative results

Complex occlusions and thin structures



Conclusions

- Unsupervised cosegmentation algorithm that uses appearance and stereo cues to:
  - infer object of interest
  - recover pixel-accurate foreground segmentation in each view
  - recover good quality depth maps

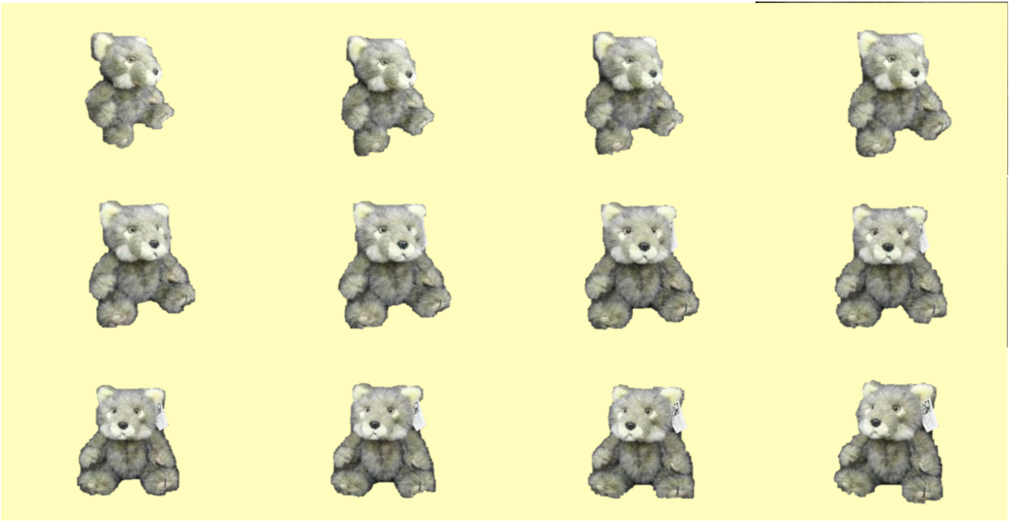
Thank you








Additional Results



Teddy sequence