CS395T paper review

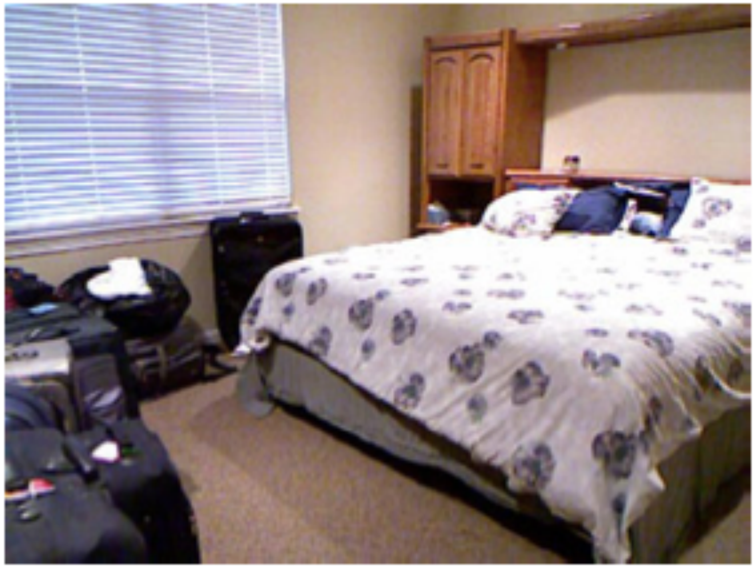
Indoor Segmentation and Support Inference from RGBD Images

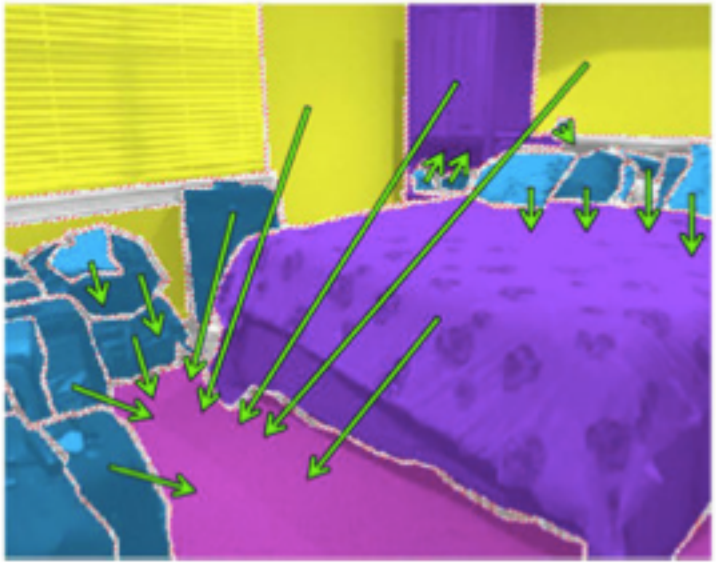
Chao Jia

Sep 28 2012

Introduction

- What do we want -- Indoor scene parsing
  - Segmentation and labeling
  - **Support relationships**





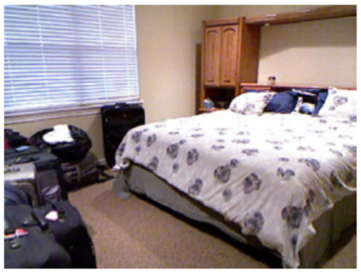
Different colors show different kinds of objects;

Support relationships help understand the scene and interact with scene elements.

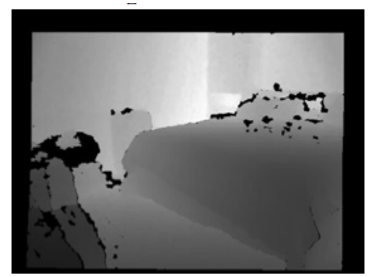
Introduction

- What do we have

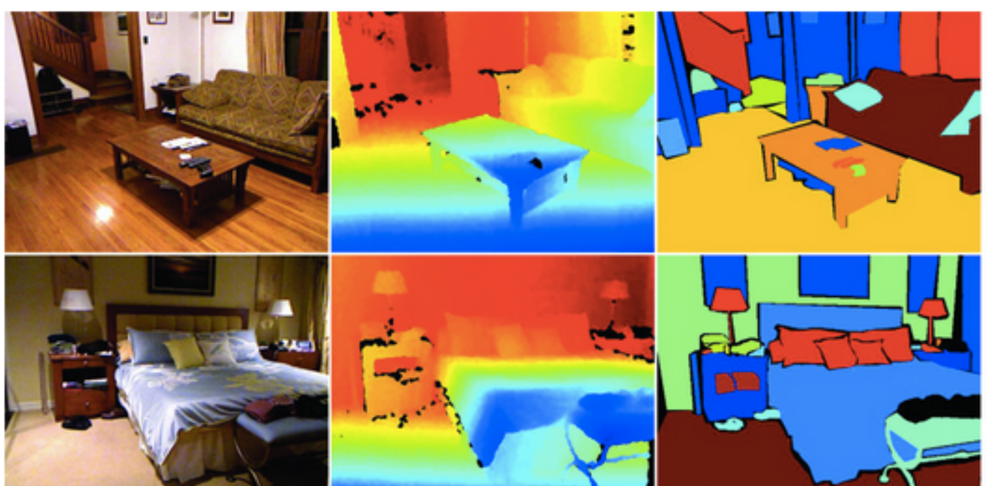
  - Color image



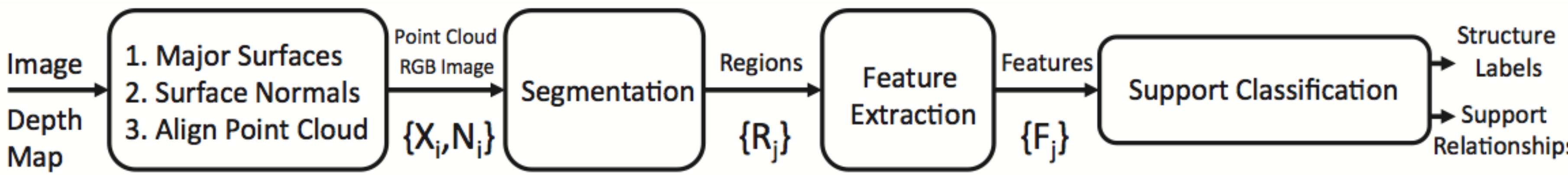
  - Depth image (3D coordinates)



- How 3D cues can best inform a structured 3D interpretation
- Dataset with 1449 densely labeled images

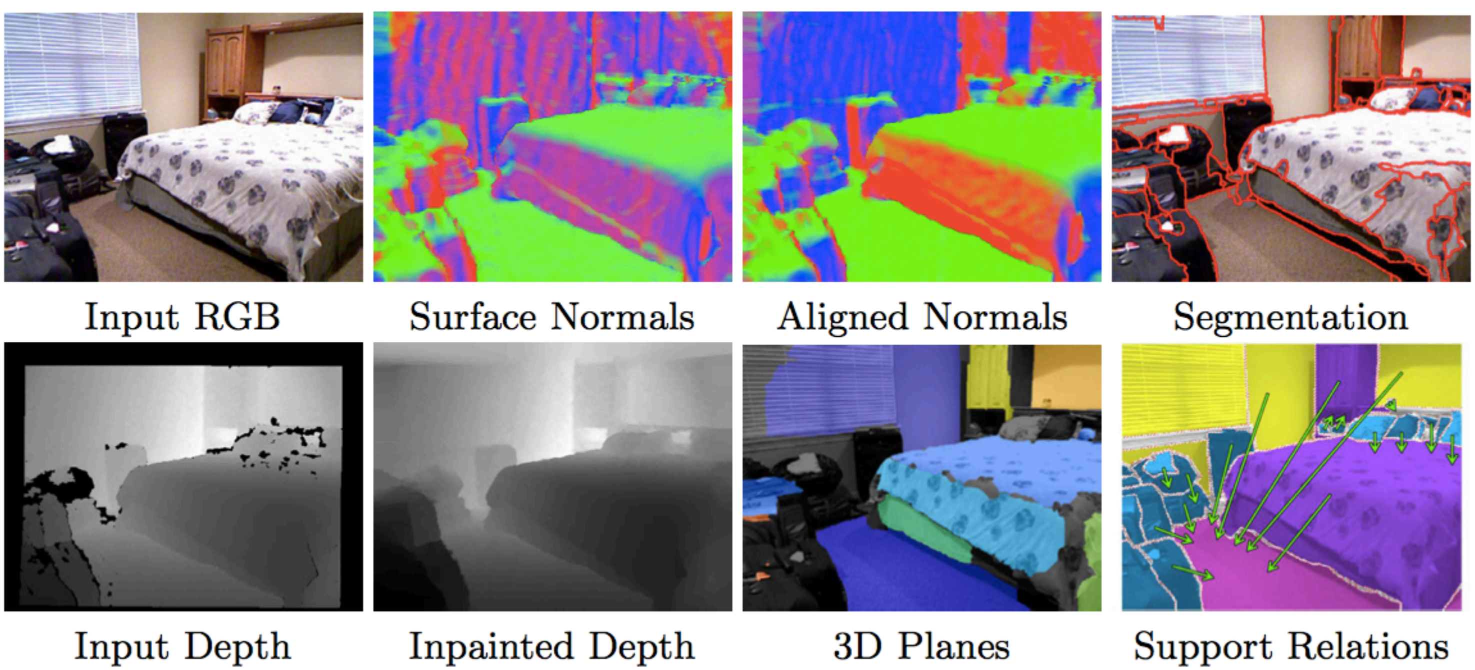


General Steps



How 3D cues help scene interpretation

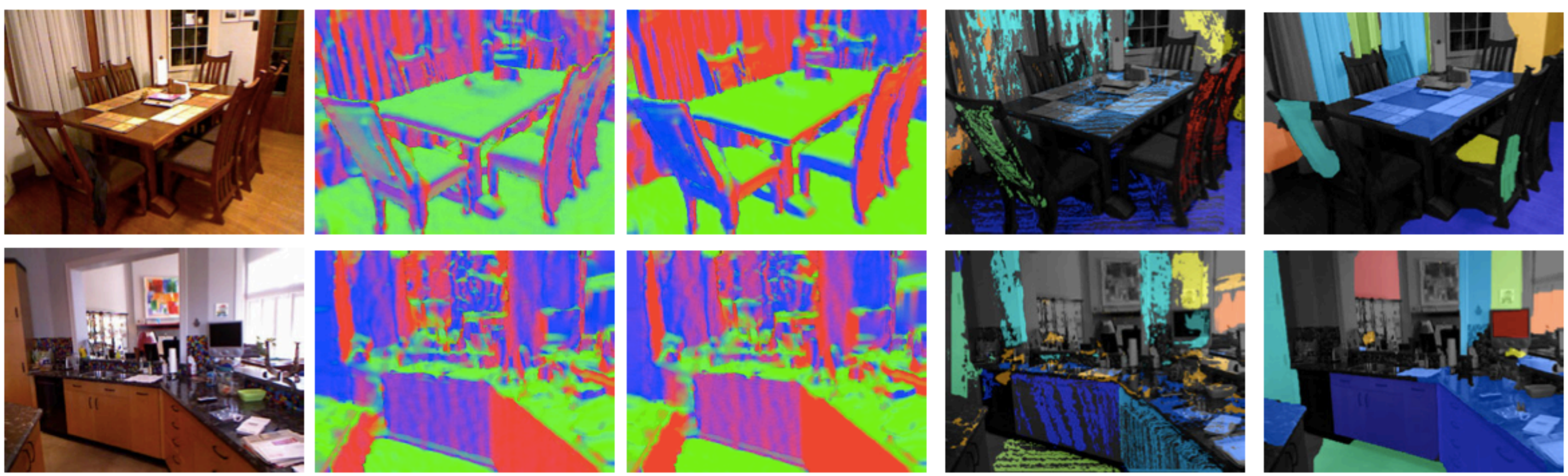
Integer programming formulation



Scene Structure Modeling

- Align the room with the 3 principle directions
  - Compute 3D lines and surface normals
  - Find the most probable X-Y-Z axis

- Segment the visible regions into 3D planes
  - Propose 3D planes using RANSAC
  - Segment the image into the proposed planes



(a) Input RGB (b) Normals (c) Aligned (d) RANSAC Fit (e) Segmented

Aligning to Room Coordinates

- Preparation using 3D coordinates
  - Straight line segments
  - 3D surface normals at each pixel
- Propose candidates (100-200)
  - All the straight 3D lines
  - Mean-shift modes of surface normals
- Search for the most probable X-Y-Z triple
  - Random sample a triple, compute the score

$$S(v_1, v_2, v_3) = \sum_{j=1}^{3} [\frac{w_N}{N_N} \sum_{i}^{N_N} \exp(\frac{-(\mathbf{N}_i \cdot \mathbf{v}_j)^2}{\sigma^2}) + \frac{w_L}{N_L} \sum_{i}^{N_L} \exp(-\frac{(\mathbf{L}_i \cdot \mathbf{v}_j)^2}{\sigma^2})]$$

  - Choose the triple with highest score
- Warp the image to align with principle directions

(a) Input RGB (b) Normals

Manhattan world assumption

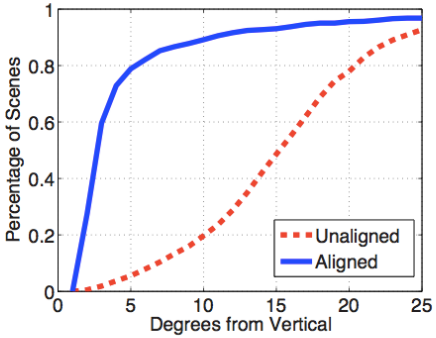


Fig. 3. Alignment of Floors

Proposing and Segmenting Planes

- Generating potential planes
  - Sample the grid of pixel and propose planes (>2500 inliers)
- Assign each pixel a label to a certain plane
  - Latent variables to infer: plane label
  - Observable variables: 3D coordinates, RGB intensities, **surface normals**
  - Conditional random field modeling solved by graph cuts

$$E(\mathbf{data}, \mathbf{y}) = \alpha_i \underbrace{\left[\sum_i f_{3d}(\mathbf{X}_i, y_i) + f_{norm}(\mathbf{N}_i, y_i) \right]}_{\text{unary term}} + \underbrace{\sum_{i,j \in \mathcal{N}_8} f_{pair}(y_i, y_j, \mathbf{I})}_{\text{pairwise term}}$$

3D coordinates

surface normals

RGB intensities

Proposing and Segmenting Planes

- Unary term $f_{3d}(\mathbf{X}_i, y_i) + f_{norm}(\mathbf{N}_i, y_i)$
  - Geometrically validate the labels

$$-\log \frac{Pr(dist|inlier)}{Pr(dist|outlier)}$$

from RANSAC plane proposing

- Pairwise term $f_{pair}(y_i, y_j, \mathbf{I})$

$$\mathbf{1}(y_i \neq y_j) \exp \underbrace{\left(-(\beta_1 + \beta_2 ||\mathbf{I}_i - \mathbf{I}_j||^2)\right)}_{\text{smoothness weighed by RGB intensity difference}}$$

Segmentation

- Oversegmentation into superpixels
  - Boundaries detection from **RGB intensities**
  - Force consistency with **3D planes regions**

- Iterative merging of regions
  - Regions with minimum boundary strength are merged
  - Boundary strength: $\mathrm{P}(y_i \neq y_j | \mathbf{x}_{ij}^s)$
    - Trained boosted decision tree classifier
    - y: labels of regions
    - x: paired regions features

Segmentation

- Paired region features
  - **RGB features**: crucial for nearby or touching objects
  - **3D features (plane labels, surface normals, depth)**: help differentiate between texture and object edges

Modeling Support Relationships

- Variables to infer for each region (R regions in total)
  - $S_i \in \{1 \dots R, h, g\}$ the support region

    supported by other regions

    supported by an invisible region

    not supported (ground)
  - $T_i \in \{0, 1\}$ supported from below/behind
  - $M_i \in \{1, 2, 3, 4\}$ structure class
    - 1: Ground
    - 2: Furniture (large objects that cannot be carried)
    - 3: Prop (small objects that can be easily carried)
    - 4: Structure (walls, ceiling, columns)

Modeling Support Relationships

- Energy minimization

$$\{\mathbf{S}^*, \mathbf{T}^*, \mathbf{M}^*\} = \arg\max_{\mathbf{S},\mathbf{T},\mathbf{M}} P(\mathbf{S}, \mathbf{T}, \mathbf{M}|I) = \arg\min_{\mathbf{S},\mathbf{T},\mathbf{M}} E(\mathbf{S}, \mathbf{T}, \mathbf{M}|I)$$

- Factorize posterior distribution

$$P(\mathbf{S}, \mathbf{T}, \mathbf{M}|I) \propto \underbrace{\prod_{i=1}^{R} P(I|S_i, T_i) P(I|M_i)}_{\text{likelihood + factorization}} \underbrace{P(\mathbf{S}, \mathbf{T}, \mathbf{M})}_{\text{Prior}}$$

- Final problem

$$E(\mathbf{S}, \mathbf{T}, \mathbf{M}) = -\sum_{i=1}^{R} \underbrace{\log(D_s(F^s_{i,S_i}|S_i, T_i) + \log(D_m(F^m_i|M_i))}_{\text{likelihood + factorization}} + \underbrace{E_P(\mathbf{S}, \mathbf{T}, \mathbf{M})}_{\text{Prior}}$$

Modeling Support Relationships

- Prior term

$$E_P(\mathbf{S},\mathbf{T},\mathbf{M}) = \sum_{i=1}^{R} \psi_{TC}(M_i, M_{S_i}, T_i) + \psi_{SC}(S_i, T_i) + \psi_{GC}(S_i, M_i) + \psi_{GGC}(\mathbf{M})$$

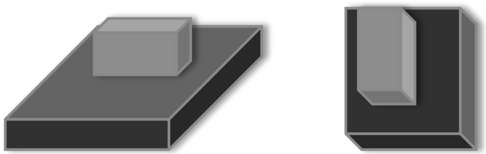
  - Transition prior (supporting relationship between two structure classes)

$$\psi_{TC}(M_i, M_{S_i}, T_i) \propto -\log \frac{\sum_{z \in supportLabels} \mathbb{1}[z = [M_i, M_{S_i}, T_i]]}{\sum_{z \in supportLabels} \mathbb{1}[z = [M_i, *, T_i]]}$$

which
combination
is more likely

  - Support consistency (between 3D structure and support relationship)

$$\psi_{SC}(S_i, T_i) = \begin{cases} (H_i^b - H_{S_i}^t)^2 & \text{if } T_i = 0 \\ V(i, S_i)^2 & \text{if } T_i = 1 \end{cases}$$



  - Global ground consistency

$$\psi_{GGC}(\mathbf{M}) = \sum_{i=1}^{R} \sum_{j=1}^{R} \begin{cases} \kappa & \text{if } M_i = 1 \ \wedge \ H_i^b > H_j^b \\ 0 & \text{otherwise,} \end{cases}$$

Everything is
above floor

  - Ground consistency

$$\psi_{GC}(S_i, M_i) = \begin{cases} \infty & \text{,if } S_i = g \text{ and } M_i \neq 1 \\ 0 & \text{,else} \end{cases}$$

No support
for floor

Modeling Support Relationships

- Likelihood term

$$-\sum_{i=1}^{R} \log(D_s(F^s_{i,S_i}|S_i,T_i)+\log(D_m(F^m_i|M_i))$$

support relation classifier   structure classifier

- F^S_{i,S_i} support features

  proximity, containment, characteristics of supporting objects, absolute 3D locations of candidate objects

- F^M_i structure features

  SIFT features, color histogram, … (object classification)

- Classifiers trained by logistic regression

Modeling Support Relationships

- Introduce Boolean indicator variables:

$$\arg\min_{\mathbf{s},\mathbf{m},\mathbf{w}} \sum_{i,j} \theta^s_{i,j} s_{i,j} + \sum_{i,u} \theta^m_{i,u} m_{i,u} + \sum_{i,j,u,v} \theta^w_{i,j,u,v} w^{u,v}_{i,j}$$

$$\begin{aligned}
\text{s.t.}\ \sum_j s_{i,j} = 1, \quad \sum_u m_{i,u} &= 1 \ \forall i \\
\sum_{j,u,v} w^{u,v}_{i,j} = 1, &\qquad \forall i \\
s_{i,2R'+1} = m_{i,1}, &\qquad \forall i \\
\sum_{u,v} w^{u,v}_{i,j} = s_{i,j}, &\qquad \forall u, v \\
\sum_{j,v} w^{u,v}_{i,j} \le m_{i,u}, &\qquad \forall i, u \\
s_{i,j},\ m_{i,u},\ w^{u,v}_{i,j} \in \{0,1\}, &\qquad \forall i, j, u, v
\end{aligned}$$

- Problem is linearized !
- Integer programming → relax the integrality constraints

Experiments

- Segmentation evaluation
  - measured as average overlap over ground truth regions for best-matching segmented region

Features	Weighted Score	Unweighted Score
RGB Only	52.5	48.7
Depth Only	55.9	47.3
RGBD	62.7	52.7
RGBD + Support	63.4	53.7
RGBD + Support + Structure classes	63.9	54.1

Support Relationships Evaluation

- Evaluate proposed inference model against
  - Image plane rules
    (no structure class assignment)
  - Structure class rules
    (class assignment by trained classifier)
  - Support classifier
    (no structure class assignment; infer the support relationship between every pair of regions)

- Metric
  - Percentage of correct supports

Support Relationships Evaluation

Predicting Support Relationships				
Region Source	Ground Truth		Segmentation	
Algorithm	Type Agnostic	Type Aware	Type Agnostic	Type Aware
Image Plane Rules	63.9	50.7	22.1	19.4
Structure Class Rules	72.0	57.7	45.8	41.4
Support Classifier	70.1	63.4	45.8	37.1
Energy Min (LP)	75.9	72.6	55.1	54.5

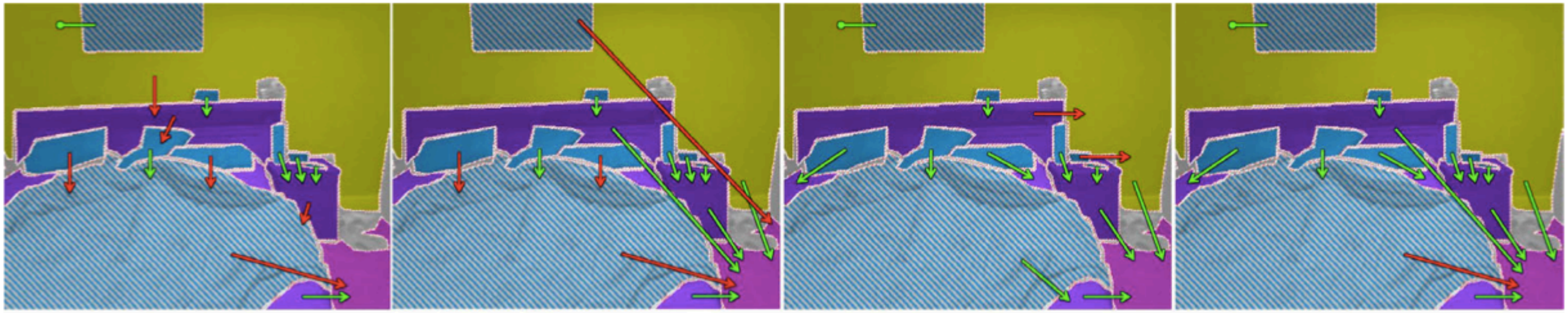


Image Plane Rules   Structure Class Rules   Support Classifier   Energy Minimization

Experiments

- Structure class prediction evaluation
  - only slightly better than local classification

Predicting Structure Classes				
	Overall		Mean Class	
Algorithm	G. T.	Seg.	G. T.	Seg.
Classifier	79.9	58.7	79.2	59.0
Energy Min (LP)	80.3	58.6	80.3	59.6

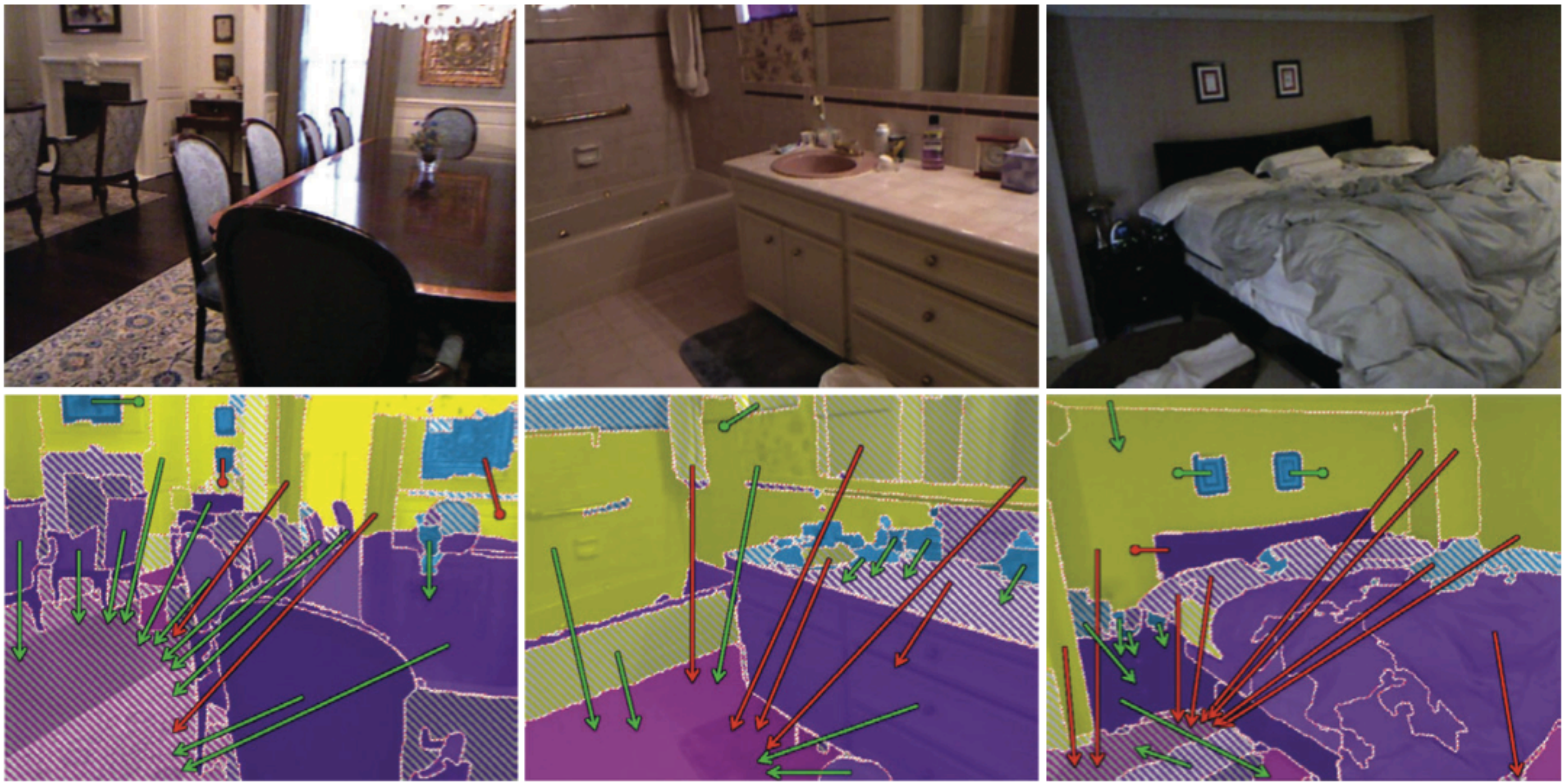
Labels \ Predictions	Ground	Furniture	Prop	Structure
Ground	.68	.28	.02	.02
Furniture	.04	.70	.14	.12
Prop	.03	.43	.42	.12
Structure	.01	.24	.14	.59

More results

- Using ground-truth segmentation

More results

- Using proposed segmentation



Summary

- Pros
  - 3D features (planes, surface normals, 3D coordinates) help segmentation and support relationship inference
  - Globally infer the support relationships with high accuracy (50% - 70%)

- Cons
  - Too many functions based on training ---- training time and training data size
  - What is a good factorization of the posterior distribution in inference of support relationships ---- Are structure class features and support features really separable ?
  - Should we consider more kinds of objects instead of just props (to make features more distinguishable) ?