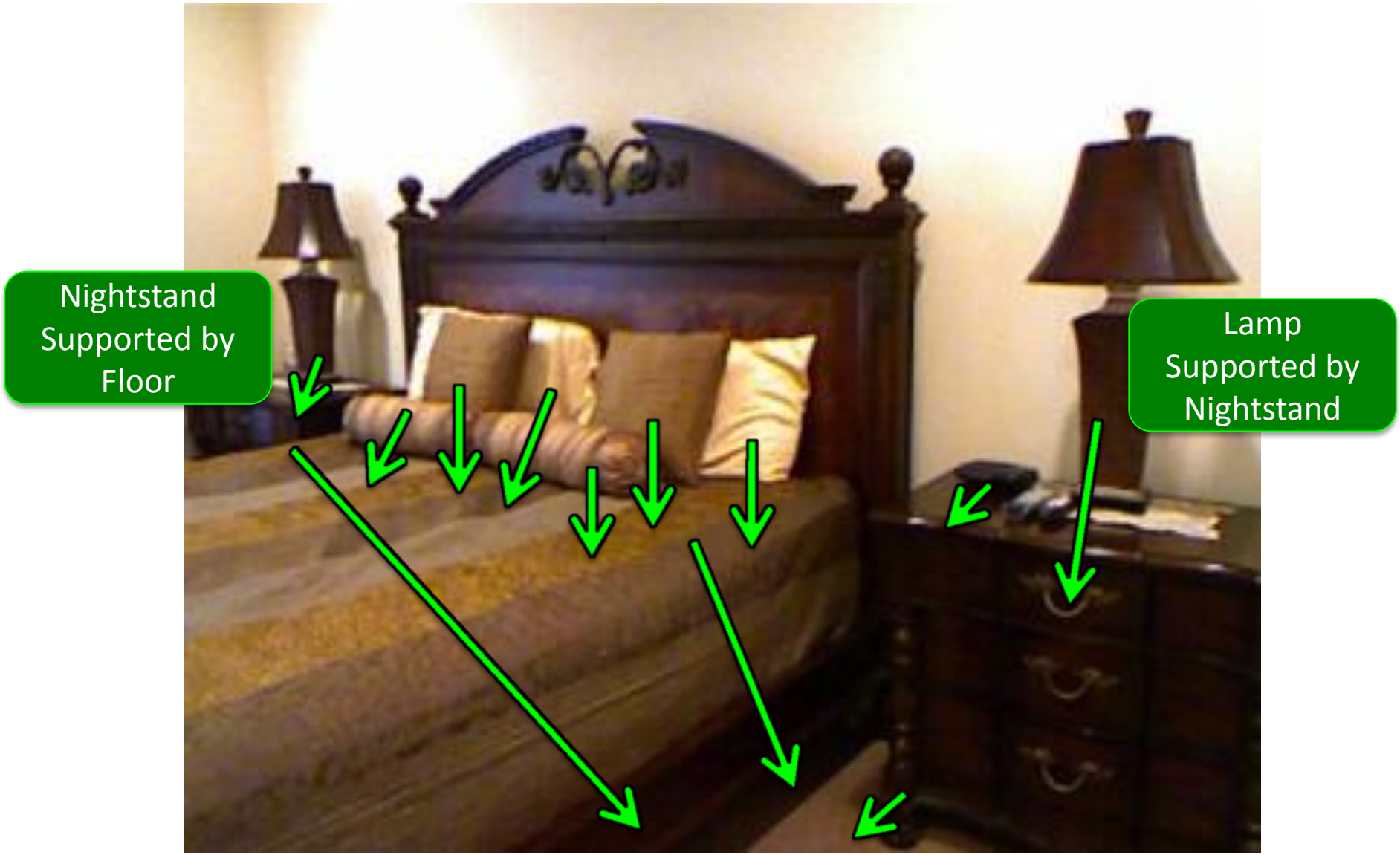
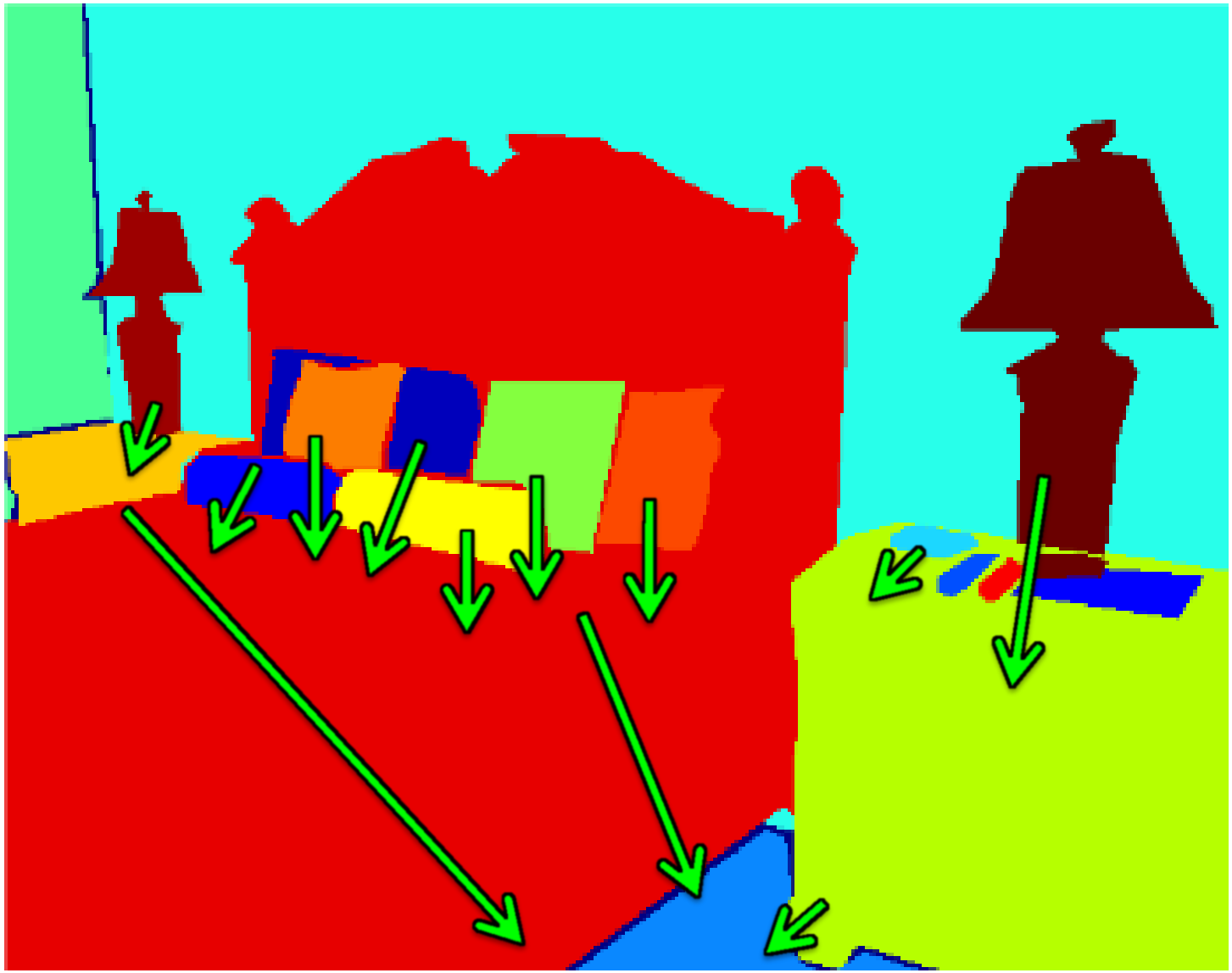
Indoor Segmentation and Support Inference from RGBD Images

Nathan Silberman, Derek Hoiem,
Pushmeet Kohli, Rob Fergus

Goal: Infer Support for Every Region



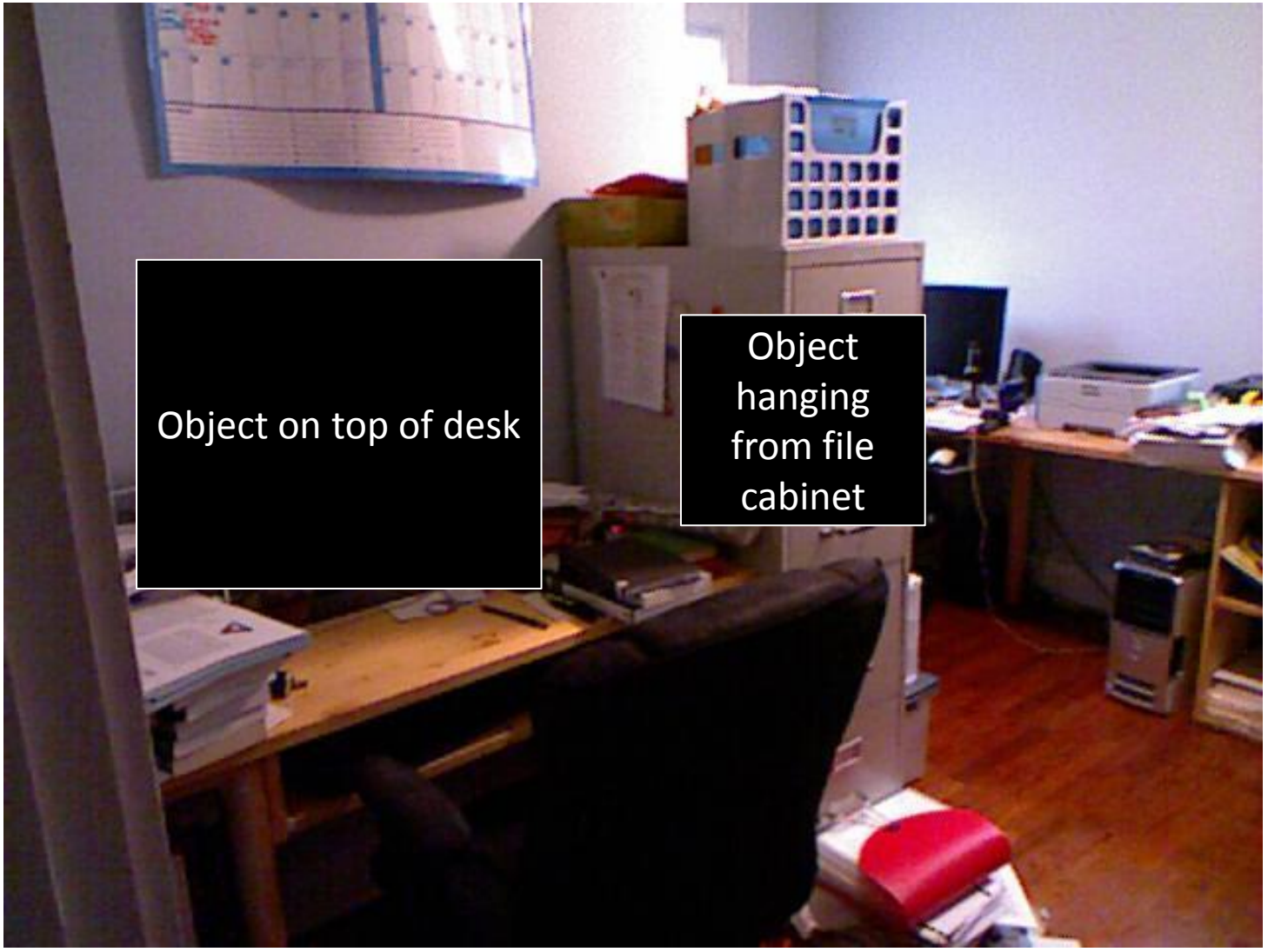
Goal: Infer Support for Every Region



Why infer physical support?

Interacting with objects may have physical consequences!

Why infer physical support: Recognition



Why infer physical support: Recognition

Working with RGB+Depth

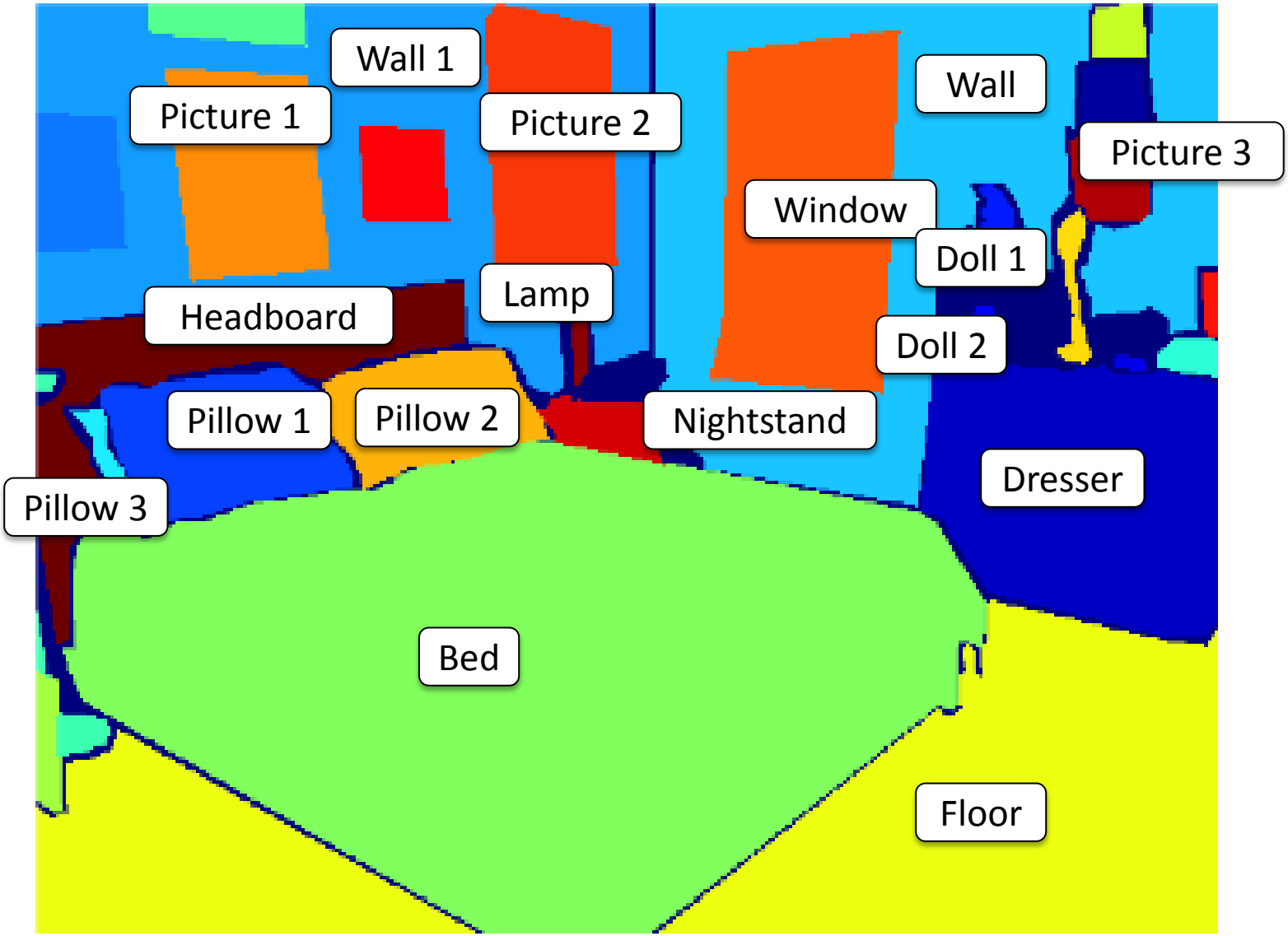
- Captured with Microsoft Kinect
- Restricted to Indoor Scenes

NYU Depth Dataset Version 2.0

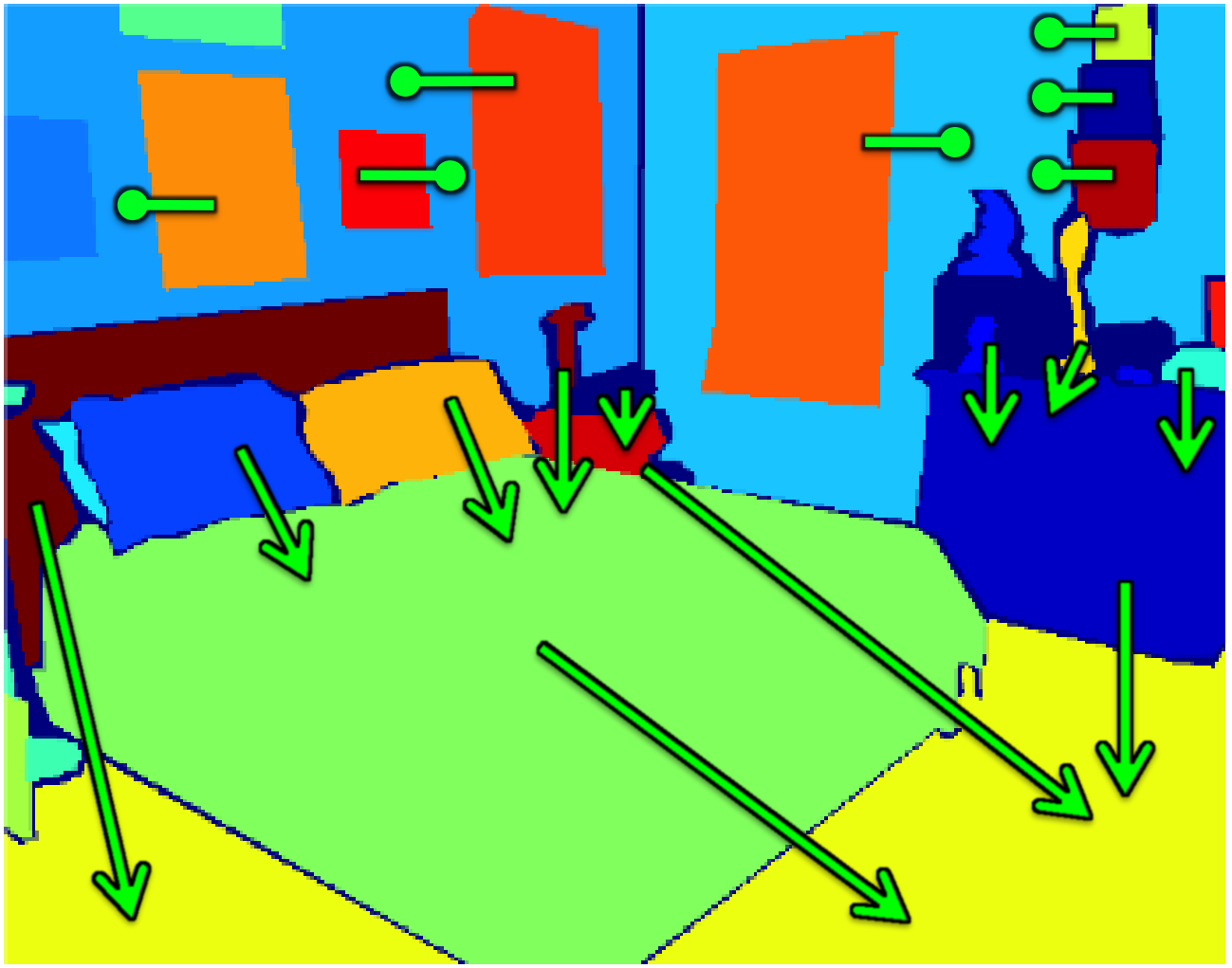
- Collected *new* NYU Depth Dataset
- Much larger than NYU Depth 1.0
  - 464 Scenes
  - 1449 Densely Labeled frames
  - Over 400,000 Unlabeled frames
  - Over 800 Semantic Classes
  - Full videos available
- Larger variation in scenes
- Dense Labels much higher quality

http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

High Quality Semantic Labels



High Quality Support Labels



Support from below

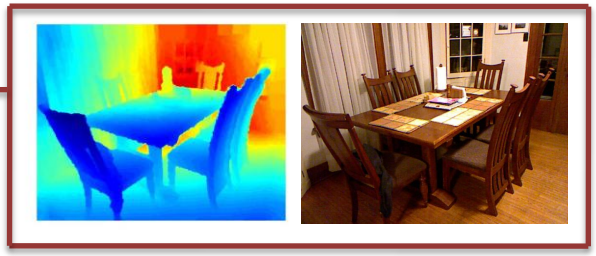
Support from behind

Support from hidden region


RGBD Image
Segmentation
Support Inference

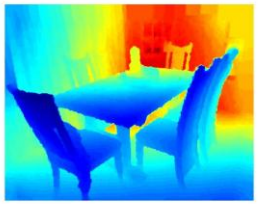
Scene Parsing

Input

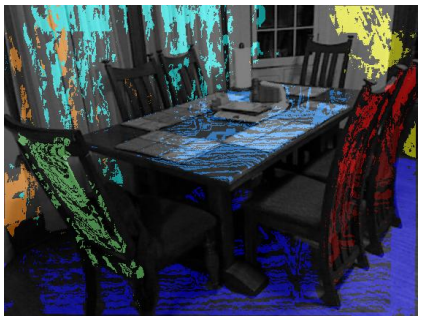


Scene Parsing

Input

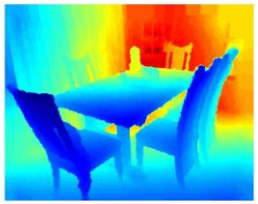


- Major Surfaces
- Surface Normals
- Aligned Point Cloud

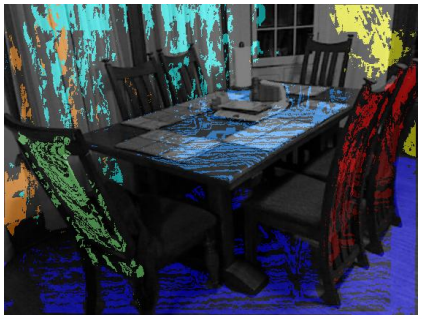


Scene Parsing

Input




- Major Surfaces
- Surface Normals
- Aligned Point Cloud



Segmentation

Hierarchical Segmentation

Segmentation Scheme similar to: **Recovering Occlusion Boundaries from a Single Image**
D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, ICCV 2007.

Hierarchical Segmentation

Segmentation Scheme similar to: **Recovering Occlusion Boundaries from a Single Image**
D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, ICCV 2007.

Hierarchical Segmentation

Segmentation Scheme similar to: **Recovering Occlusion Boundaries from a Single Image**
D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, ICCV 2007.

Hierarchical Segmentation

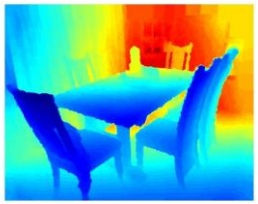
Segmentation Scheme similar to: **Recovering Occlusion Boundaries from a Single Image**
D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, ICCV 2007.

Hierarchical Segmentation

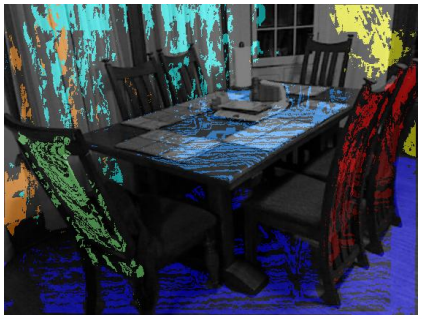
Segmentation Scheme similar to: **Recovering Occlusion Boundaries from a Single Image**
D. Hoiem, A.N. Stein, A.A. Efros, and M. Hebert, ICCV 2007.

Scene Parsing

Input



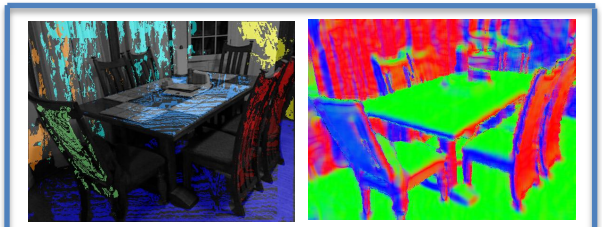

- Major Surfaces
- Surface Normals
- Aligned Point Cloud



Segmentation

Scene Parsing

Input



- Major Surfaces
- Surface Normals
- Aligned Point Cloud

Segmentation

Support Classification


RGBD Image
Segmentation
Support Inference

Modeling Choice #1

- All objects supported by a single object except –
- Floor requires no support.

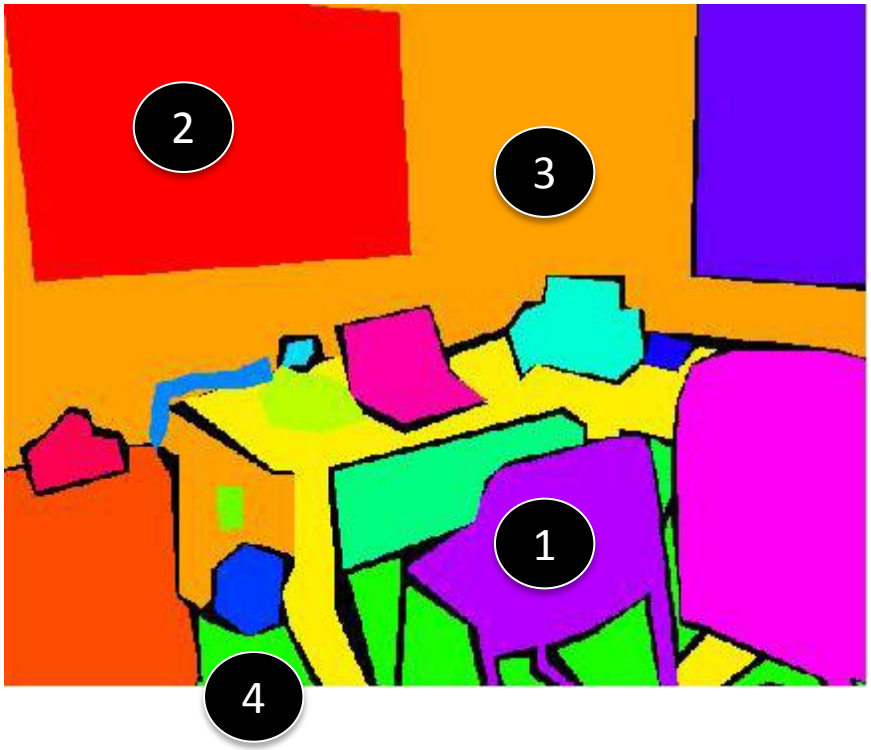
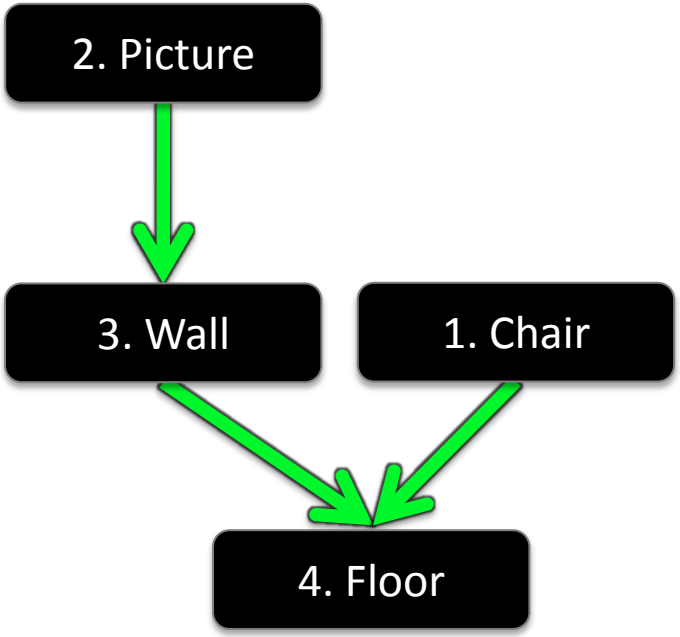


Image Regions



(Inverted) Tree Representation

Modeling Choice #2

All objects are either supported by another region in the image OR a hidden region.

Modeling Choice #2

All objects are either supported by another region in the image OR a hidden region.

Modeling Choice #2

All objects are either supported by another region in the image OR a hidden region.

Modeling Choice #3

Every object is either supported from *below* or from *behind*.

Modeling Choice #3

Every object is either supported from *below* or from *behind*.

Modeling Choice #3

Every object is either supported from *below* or from *behind*.

Modeling Support: Structure Classes

‘Structure Classes’ encode high level support
prior knowledge

(1) Ground (2) Furniture (3) Prop or (4) Structure

Modeling Support

Goal: For each region i in R regions, infer:

1. Supporting region $S_i \in \{1..R, hidden, \emptyset\}$
2. Support Type $T_i \in \{below, behind\}$
3. Structure class

$$M_i \in \{floor, furniture, prop, structure\}$$

Modeling Support

Goal: For each region i in R regions, infer:

1. Supporting region $S_i \in \{1..R, hidden, \emptyset\}$
2. Support Type $T_i \in \{below, behind\}$
3. Structure class

$$M_i \in \{floor, furniture, prop, structure\}$$

The formal problem per image:

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

$$= \arg\min_{S,T,M} E(S,T,M|I)$$

Joint Energy Factorizes into three terms:

$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

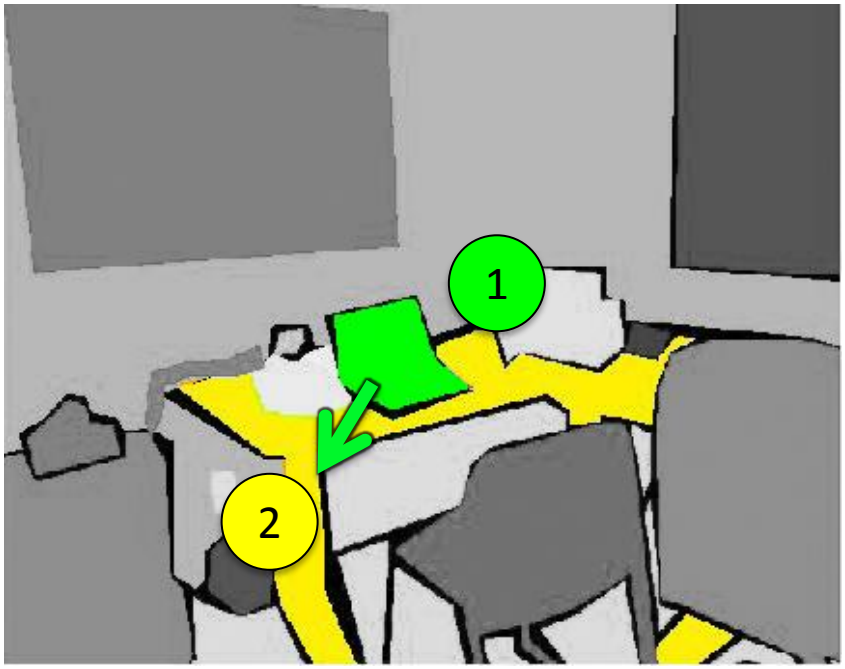
Local Support | Local Structure Class | Prior

S_i - supporting region T_i - support type M_i - structure class

$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

Local Support Energy

S_i - supporting region T_i - support type M_i - structure class



$P(S_i, T_i)$ comes from logistic regressor trained on pairwise features

$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

Local Structure Class Energy

S_i - supporting region T_i - support type M_i - structure class

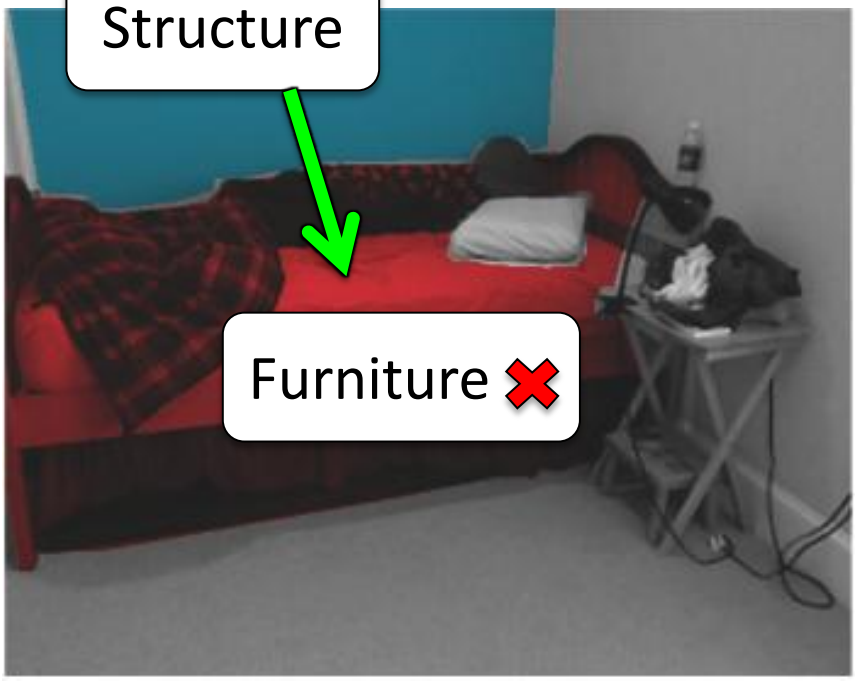
$P(M_i)$ from logistic regressor trained on features from each individual region

$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

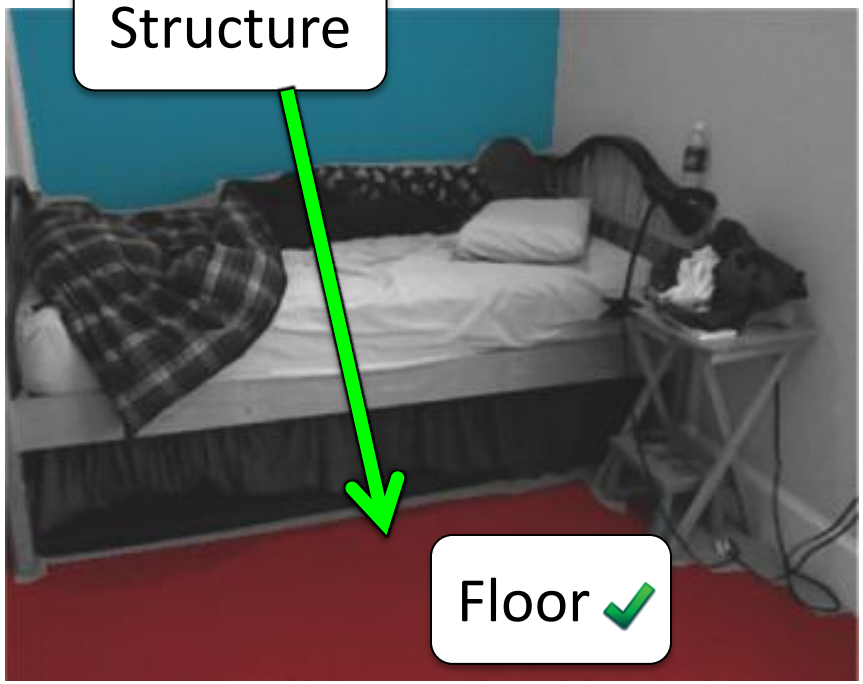
Prior (1/4): Transitions

S_i - supporting region T_i - support type M_i - structure class

A region's structure class helps predict its support.



OR



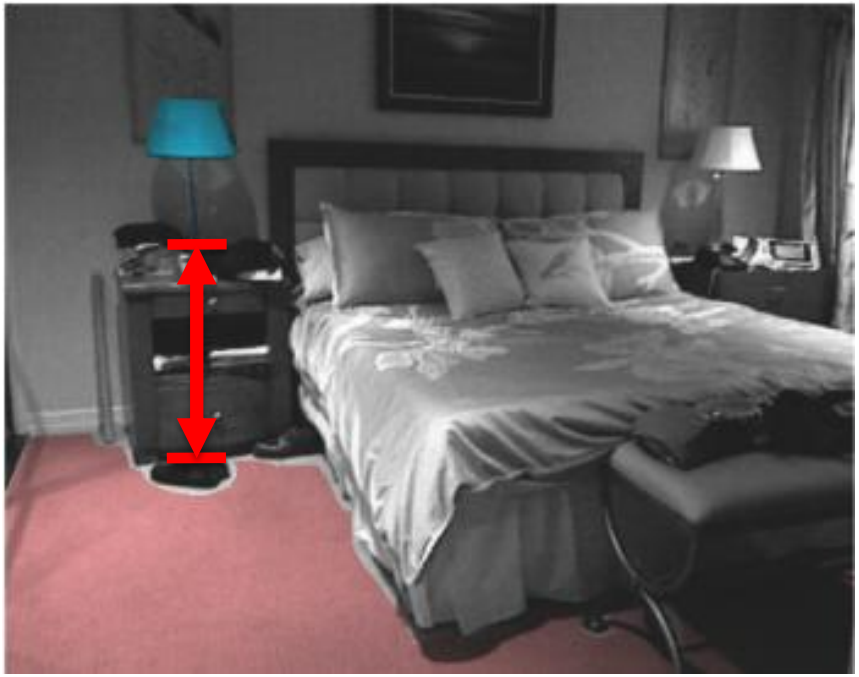
$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

Prior (2/4): Support Consistency

S_i - supporting region T_i - support type M_i - structure class

Supporting regions should be nearby

OR



$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

Prior (3/4): Ground Consistency

S_i - supporting region T_i - support type M_i - structure class

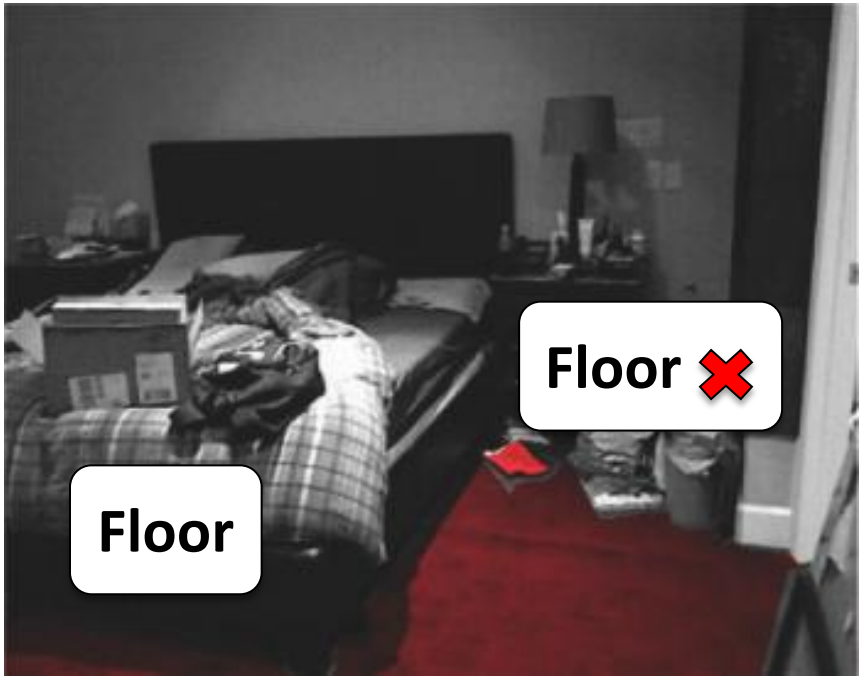
A region requires no support **if and only if** its structure class is ‘floor’

$$E(S,T,M) = \sum_{i=1}^{R} E_S(S,T) + E_M(M) + E_P(S,T,M)$$

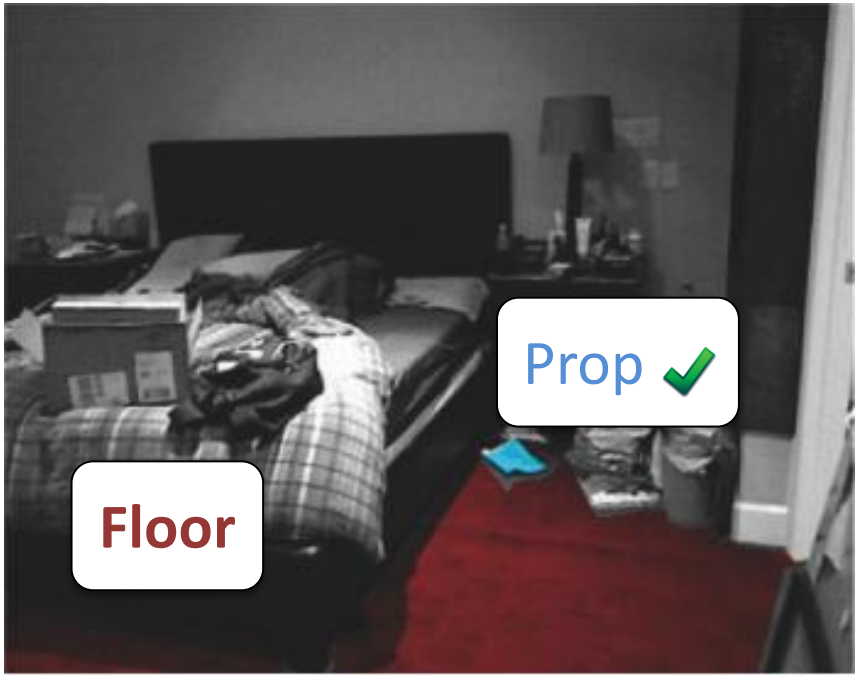
Prior (4/4): Global Ground Consistency

S_i - supporting region T_i - support type M_i - structure class

A region is unlikely to be the floor if another floor region is lower than it



OR



Integer Program Formulation

$$= \operatorname*{argmin}_{s,t,m} \sum_{i,j} \theta^s_{i,j} s_{i,j} + \sum_{i,u} \theta^m_{i,u} m_{i,u} + \sum_{i,j,u,v} \theta^w_{i,j,u,v} w^{u,v}_{i,j}$$

$$s.t. \sum_j s_{i,j} = 1, \sum_u m_{i,u} = 1 \; \forall i$$

$$\sum_{j,u,v} w^{u,v}_{i,j} = 1 \; \forall i$$

$$s_{i,2R'+1} = m_{i,1} \; \forall i$$

$$\sum_{u,v} w^{u,v}_{i,j} = s_{i,j} \; \forall u, v$$

$$\sum_{j,v} w^{u,v}_{i,j} \leq m_{i,u} \; \forall i, u$$

$$s_{i,j}, m_{i,u} w^{u,v}_{i,j} \in \{0, 1\}, \; \forall i, j, u, v$$

Relaxed to Linear Program

Experiments

Evaluating Support

$$\text{Accuracy} = \frac{\#\text{ of Correctly Labeled Support Relationships}}{\#\text{ of Total Labeled Support Relationships}}$$

Evaluation with features extracted from:

Regions from Ground Truth Labels

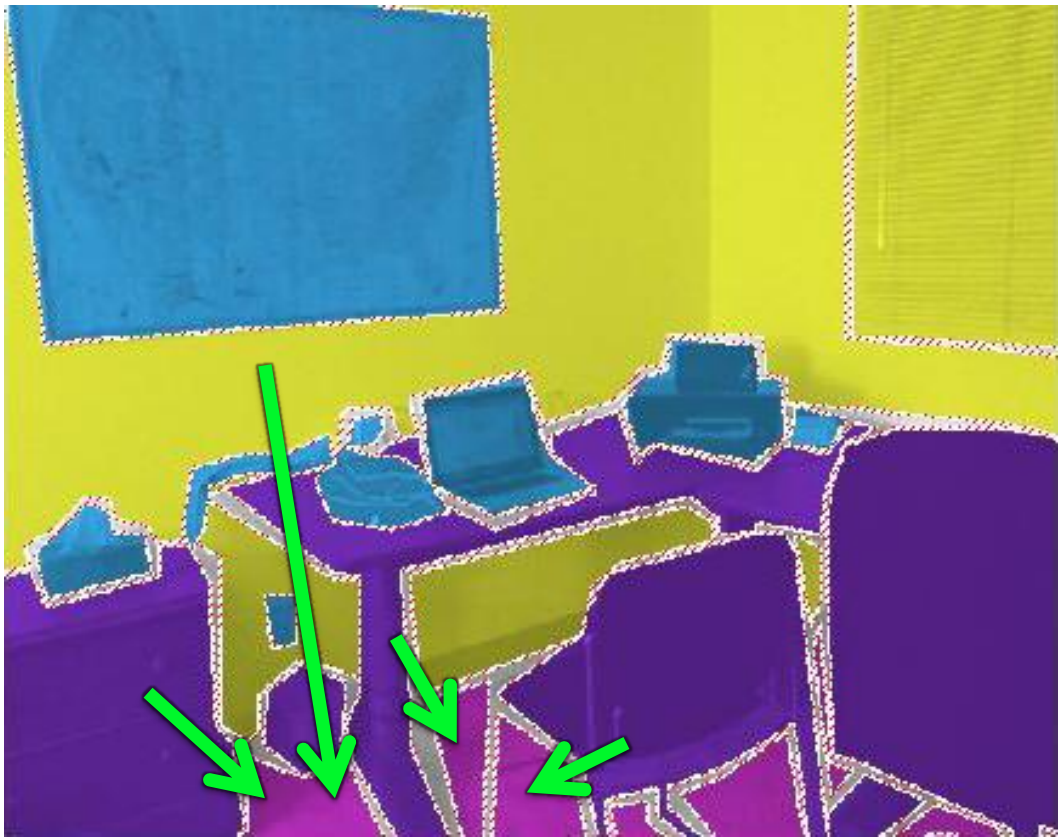
Regions from Segmentation

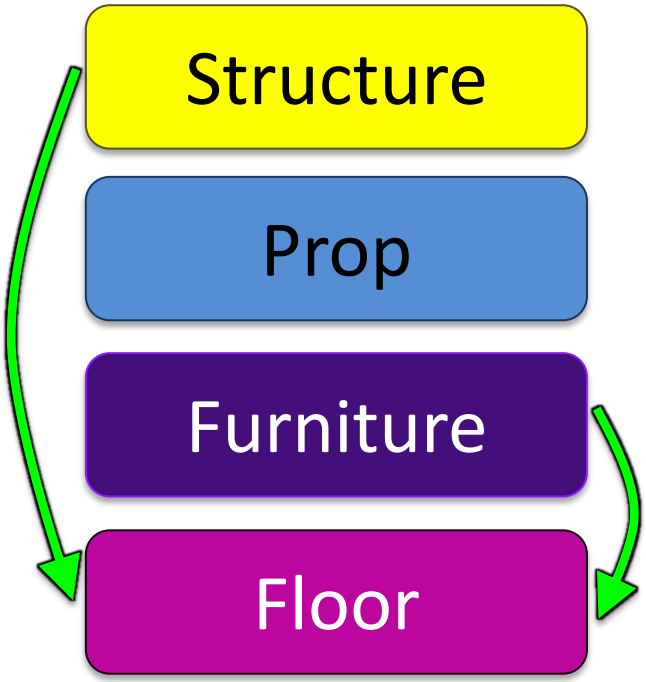
Baseline #1: Image Plane Rules

- Heuristic: look at neighboring regions for support

Baselines #2: Structure Class Rules

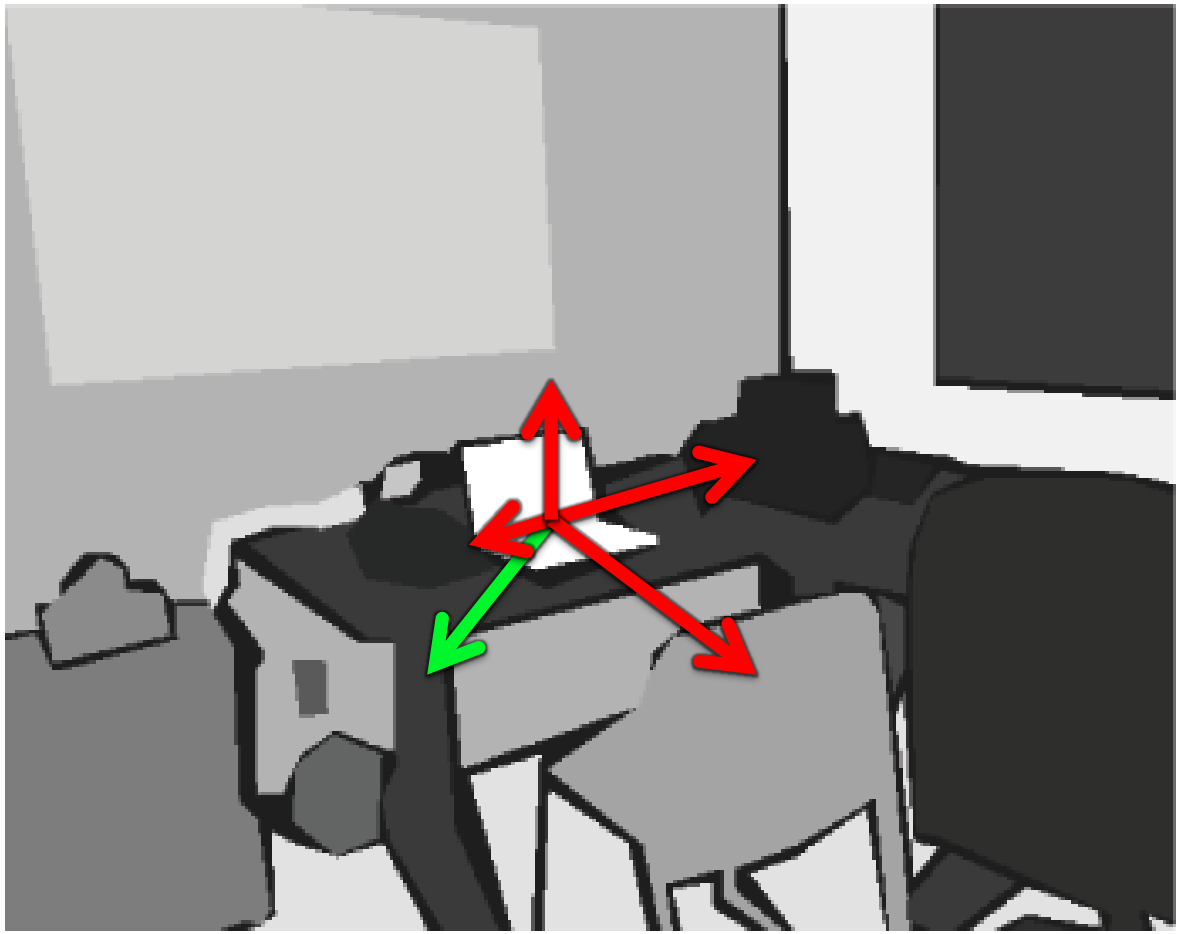
- Heuristic: Support is deterministic given Structure Classes





Baselines #3: Support Classifier

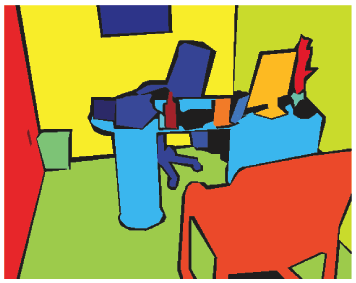
- Use only the output of support classifier

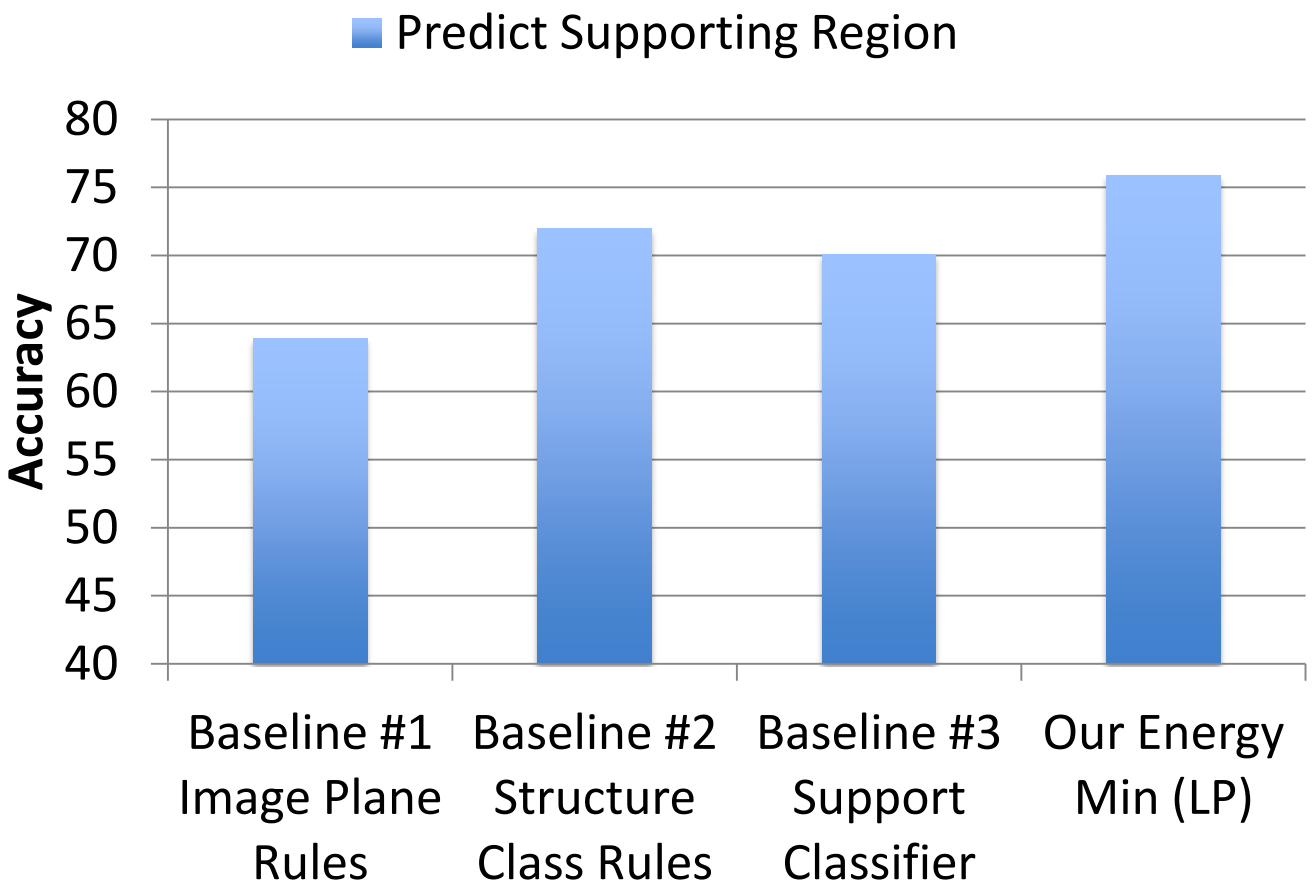


Evaluating Support

(Regions from Ground Truth Labels)

Examples of Manually Labeled Regions

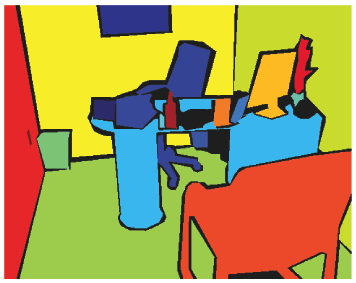


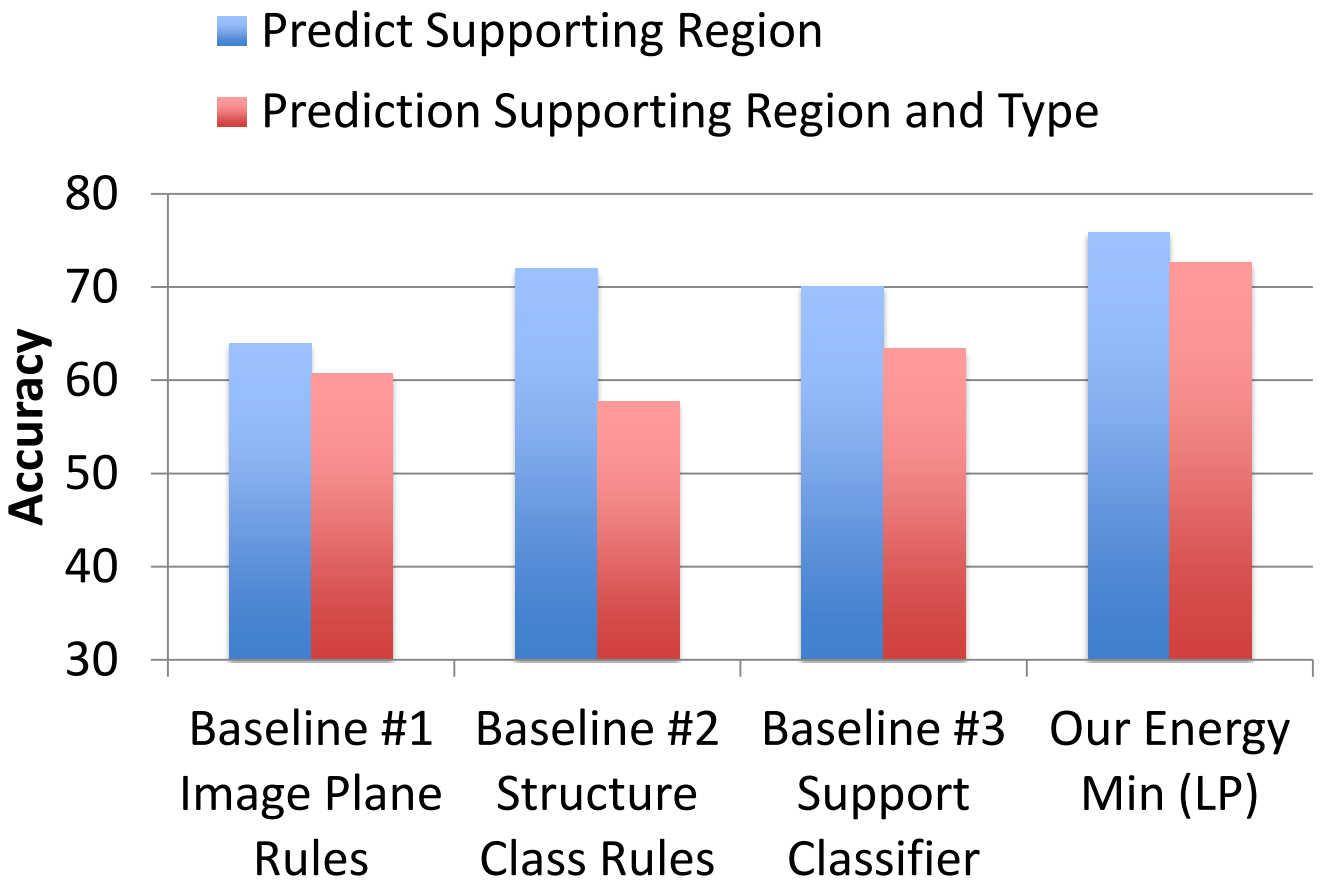


Evaluating Support

(Regions from Ground Truth Labels)

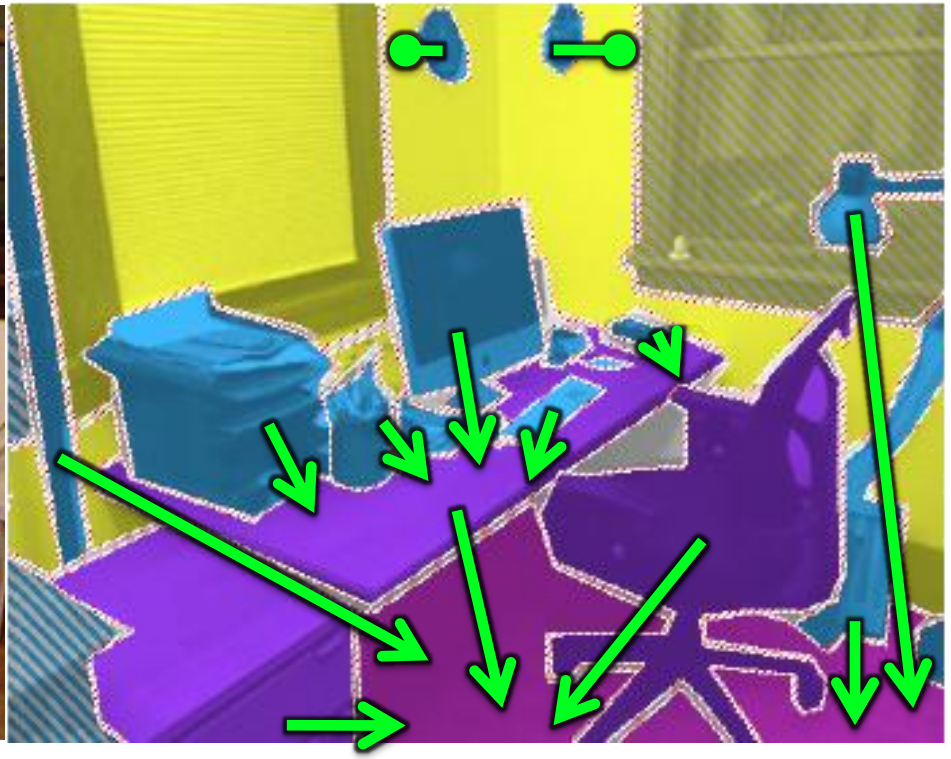
Examples of Manually Labeled Regions





Results

Ground Truth Regions



Floor

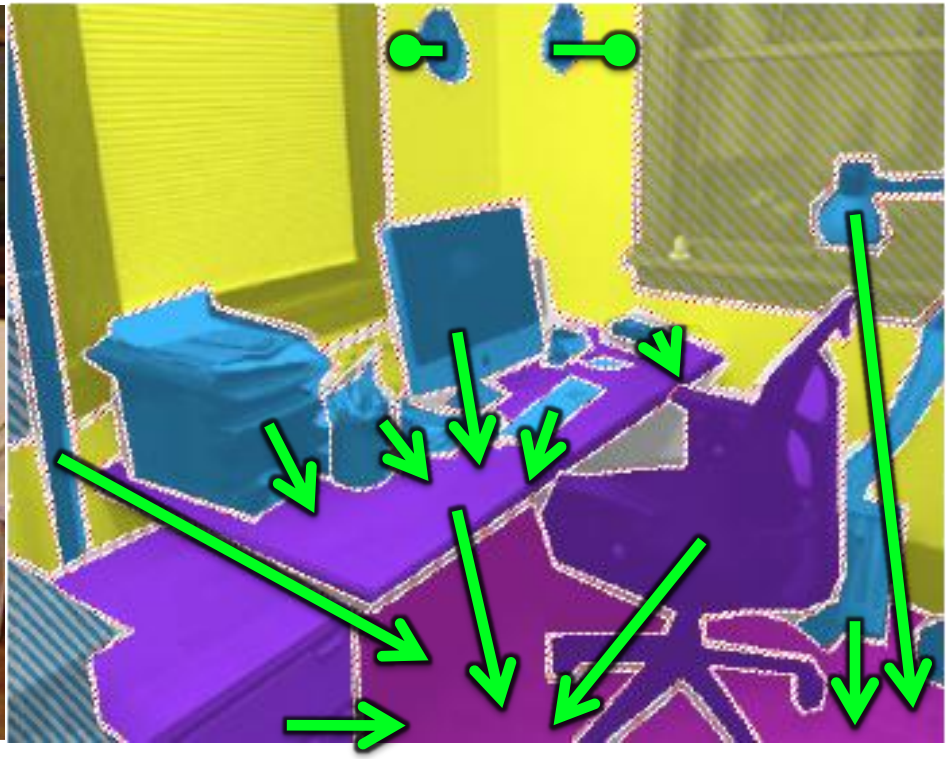
Prop

Furniture

Structure

Results

Ground Truth Regions



Correct Prediction

Results

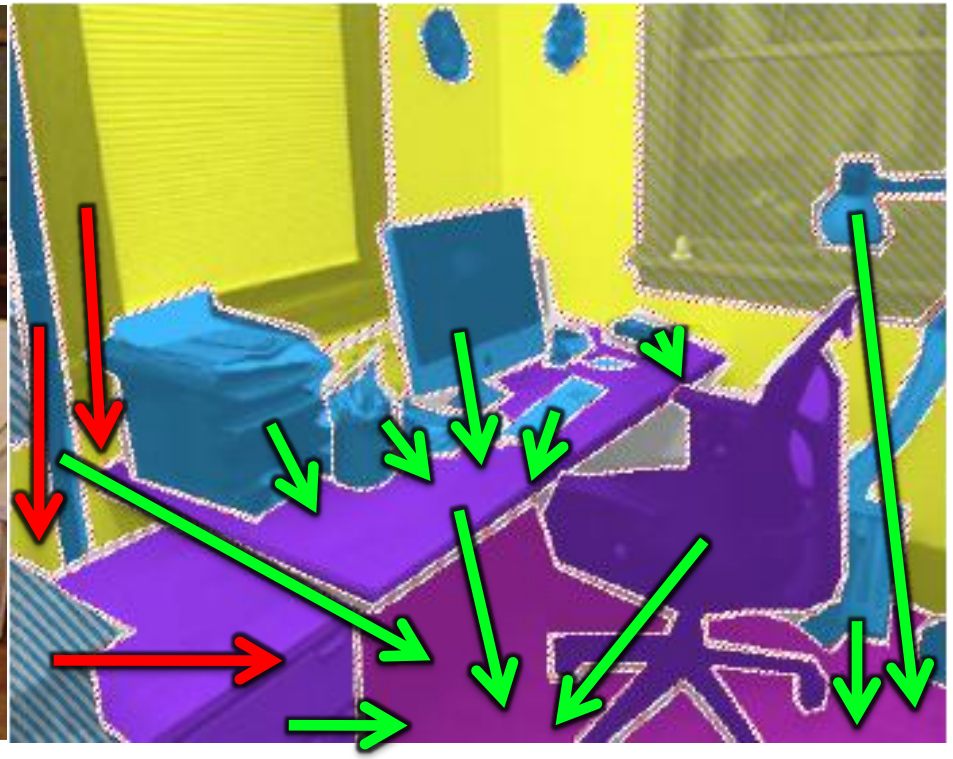
Ground Truth Regions

Correct Prediction

Incorrect Prediction

Results

Ground Truth Regions



Support from below

Correct Prediction

Incorrect Prediction

Results

Ground Truth Regions

Correct Prediction

Incorrect Prediction

Support from below

Support from behind

Results

Ground Truth Regions

Correct Prediction

Incorrect Prediction

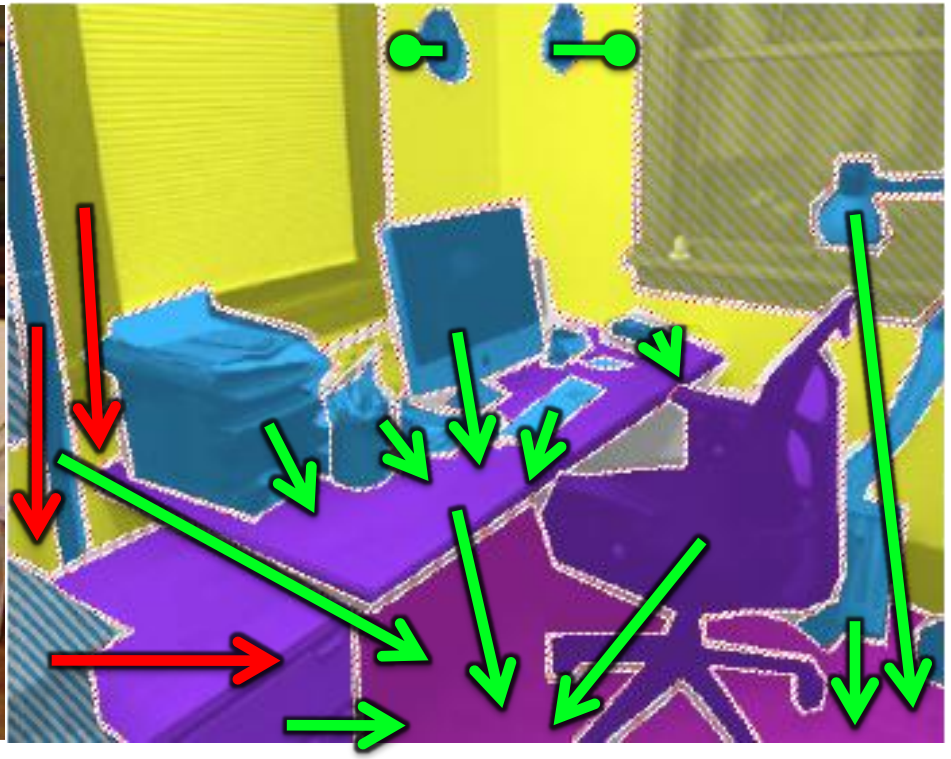
Support from below

Support from behind

Support from hidden region

Results

Ground Truth Regions



Correct Prediction

Incorrect Prediction

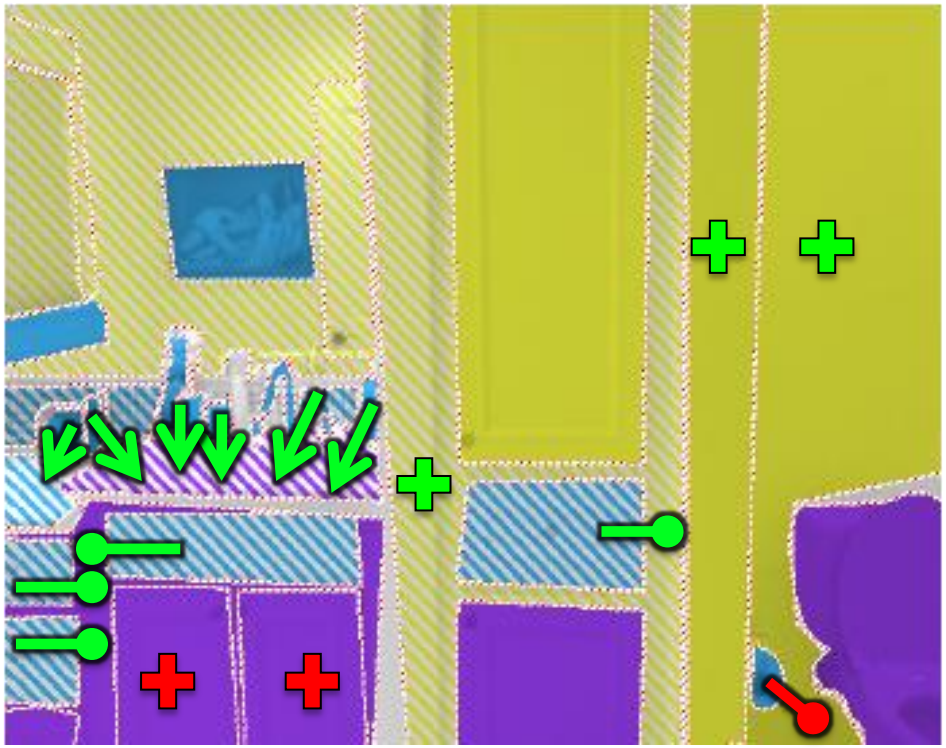
Support from below

Support from behind

Support from hidden region

Results

Ground Truth Regions



Correct Prediction

Incorrect Prediction

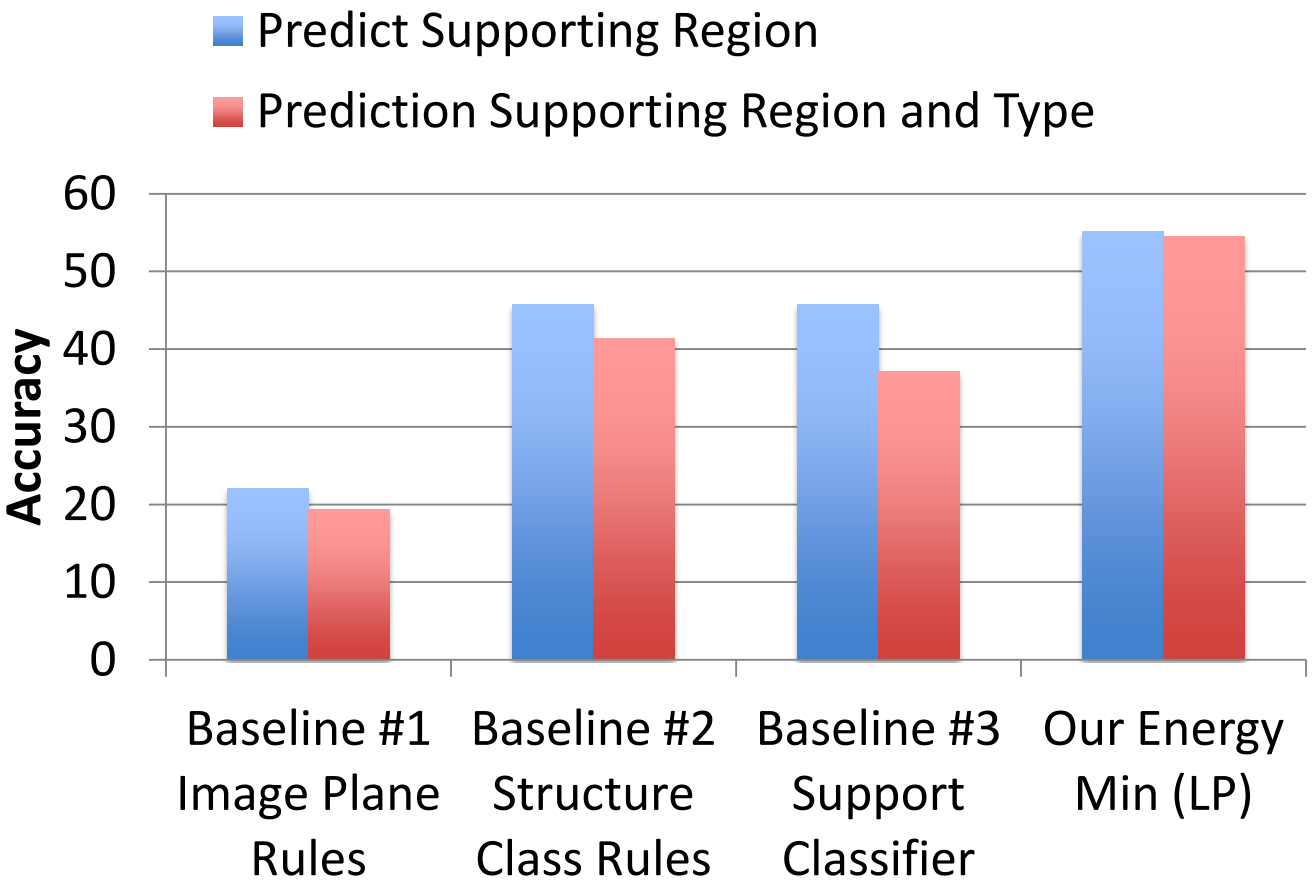
Support from below

Support from behind

Support from hidden region

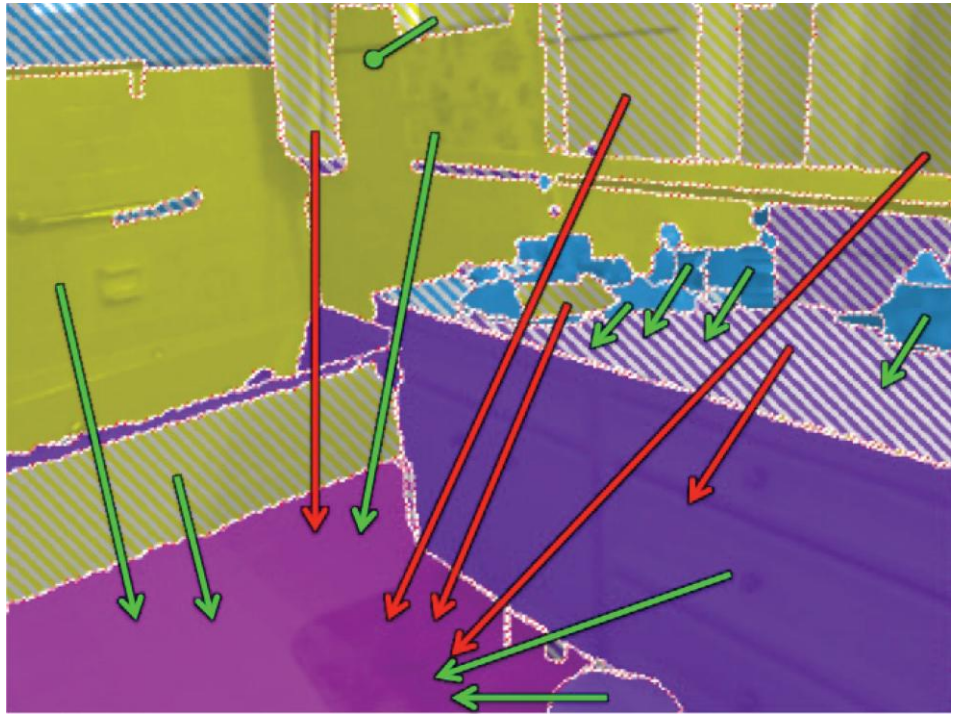
Evaluating Support
(Regions from Segmentation)

Examples of Regions from Segmentation



Results

Automatically Segmented Regions



Correct Prediction

Incorrect Prediction

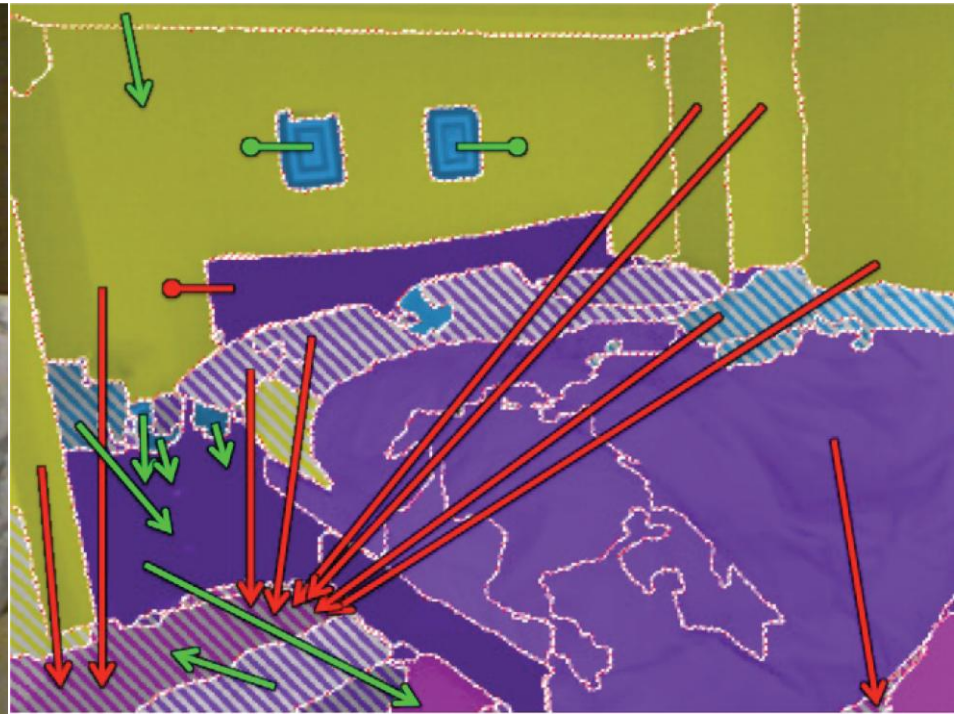
Support from below

Support from behind

Support from hidden region

Results

Automatically Segmented Regions



Correct Prediction

Incorrect Prediction

Support from below

Support from behind

Support from hidden region

Conclusion

- Algorithm for inferring Physical Support
- Novel Integer Program Formulation
- 3D Cues for segmentation

Dataset:

- http://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html

Code:

- http://cs.nyu.edu/~silberman/projects/indoor_scene_seg_sup.html