Learning Semantic Environment Perception for Cognitive Robots

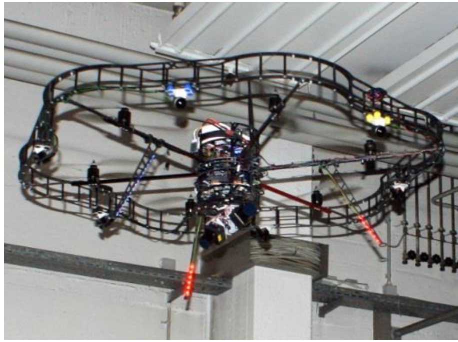
Sven Behnke

University of Bonn, Germany

Computer Science Institute VI

Some of Our Cognitive Robots

- Equipped with many sensors and DoFs
- Demonstration in complex scenarios



MAV

Soccer robot

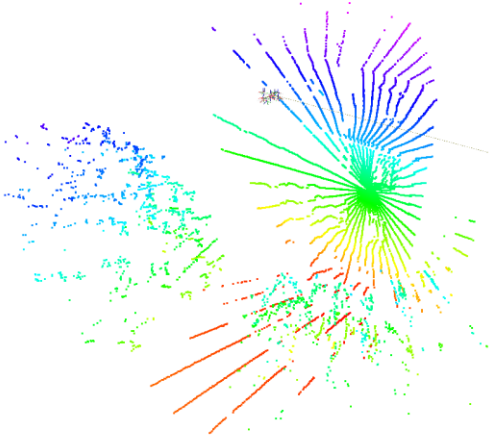
Service robot

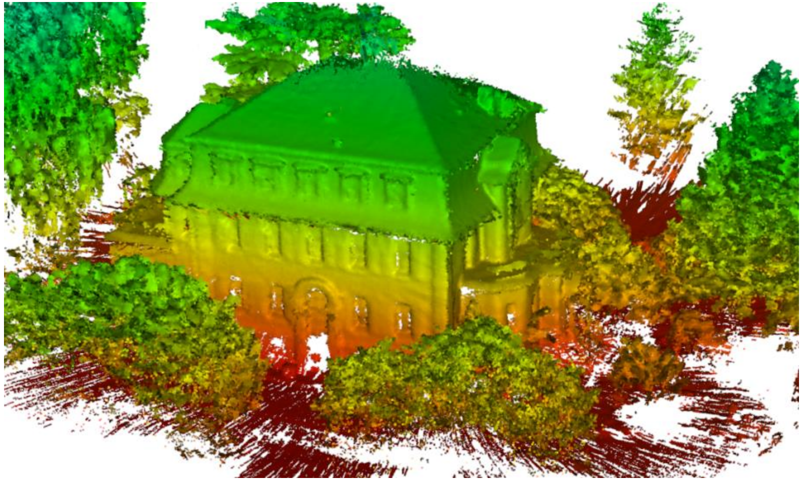
Exploration robot

Picking robot

3D Environment Perception

- 3D laser scanner, dual wide-angle stereo cameras, ultrasound, Quad Core i7
- Autonomous navigation close to structures
- 3D mapping

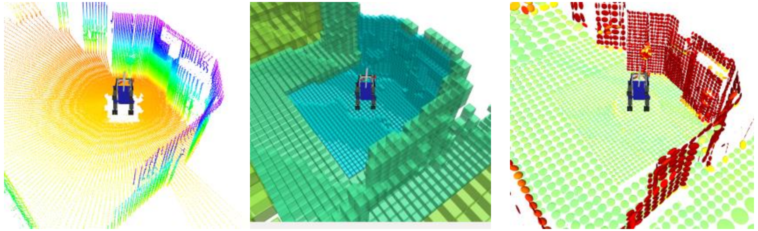


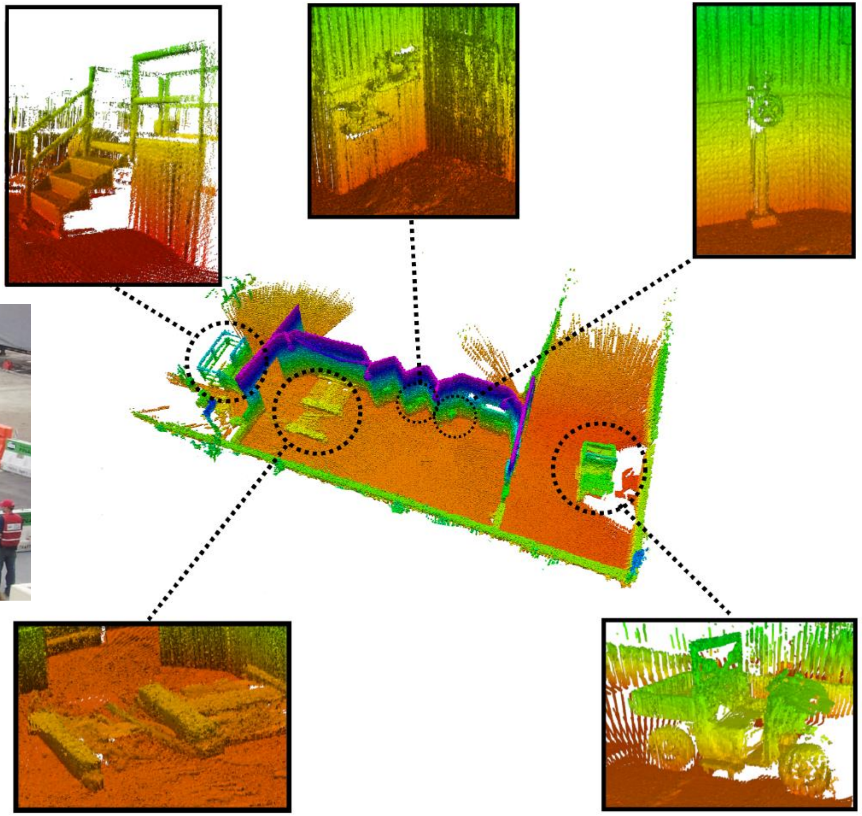


[Droeschel et al. JFR 2016]

3D Mapping

- Registering 3D laser scans



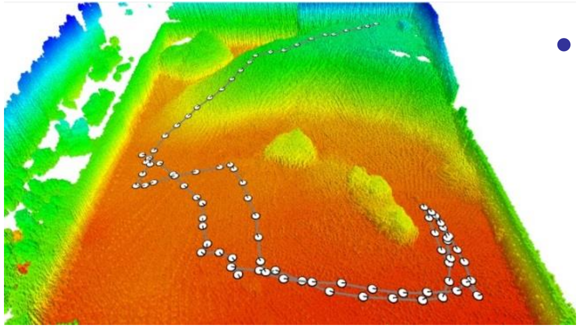


[Droeschel et al. 2016]

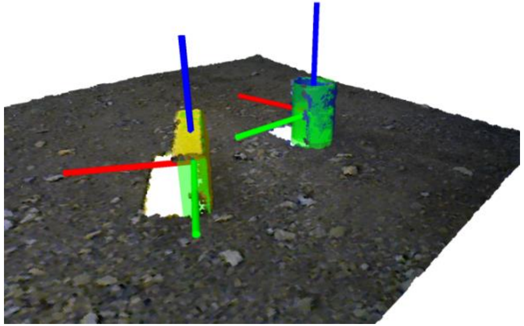
Mobile Manipulation in Mars-like Environment

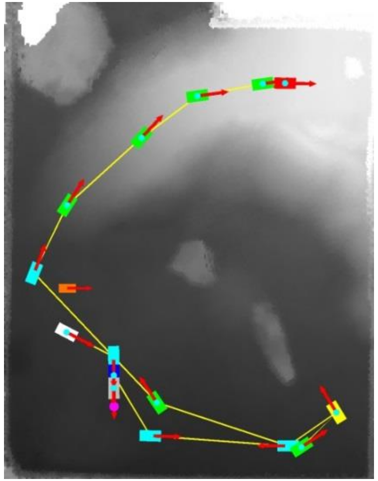
Autonomous Mission Execution

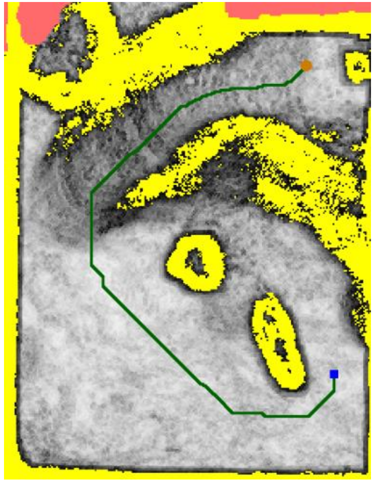
- 3D mapping, localization, mission and navigation planning

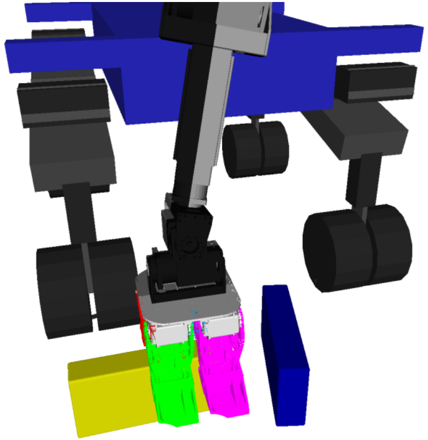


- 3D object perception and grasping









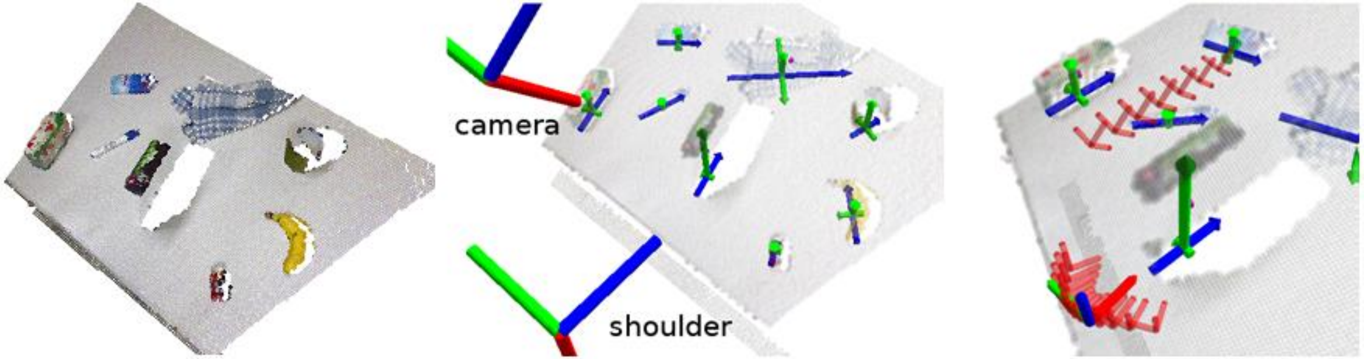
[Schwarz et al. Frontiers 2016]

universitätbonn ais

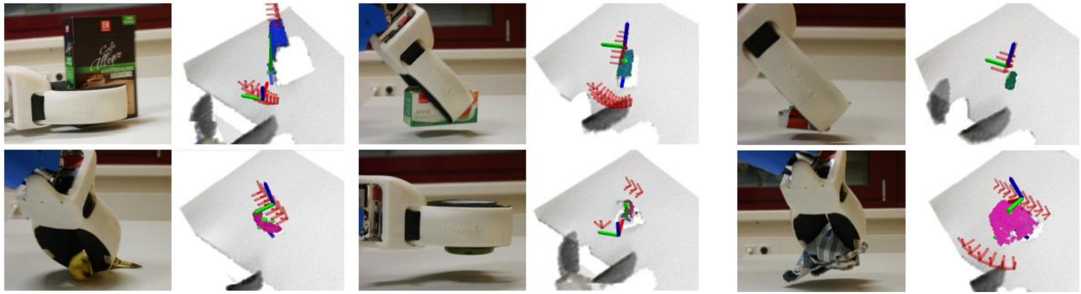
Cognitive Service Robot Cosero

Table-top Analysis and Grasp Planning

- Detection of clusters above horizontal plane
- Two grasps (top, side)



- Flexible grasping of many unknown objects

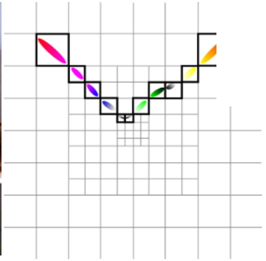


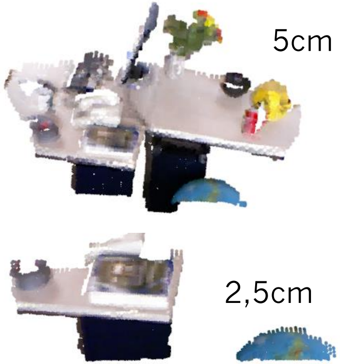
[Stückler et al, Robotics and Autonomous Systems, 2013]

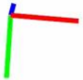
3D Mapping by RGB-D SLAM

[Stückler, Behnke: Journal of Visual Communication and Image Representation 2013]

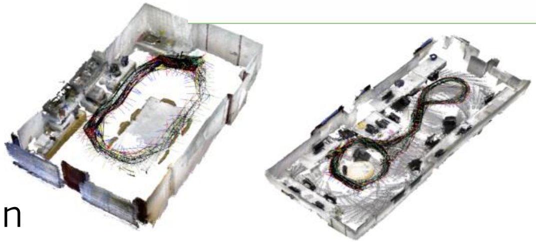
- Modelling of shape and color distributions in voxels
- Local multiresolution
- Efficient registration of views on CPU
- Global optimization







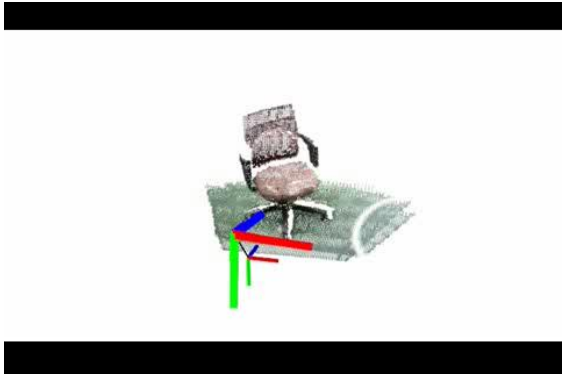
- Multi-camera SLAM



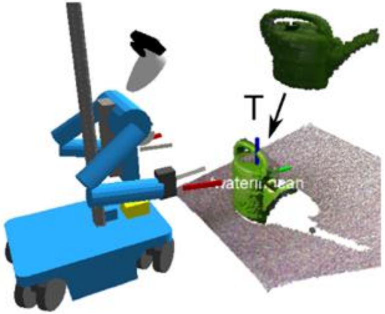
[Stoucken]

Learning and Tracking Object Models

- Modeling of objects by RGB-D-SLAM

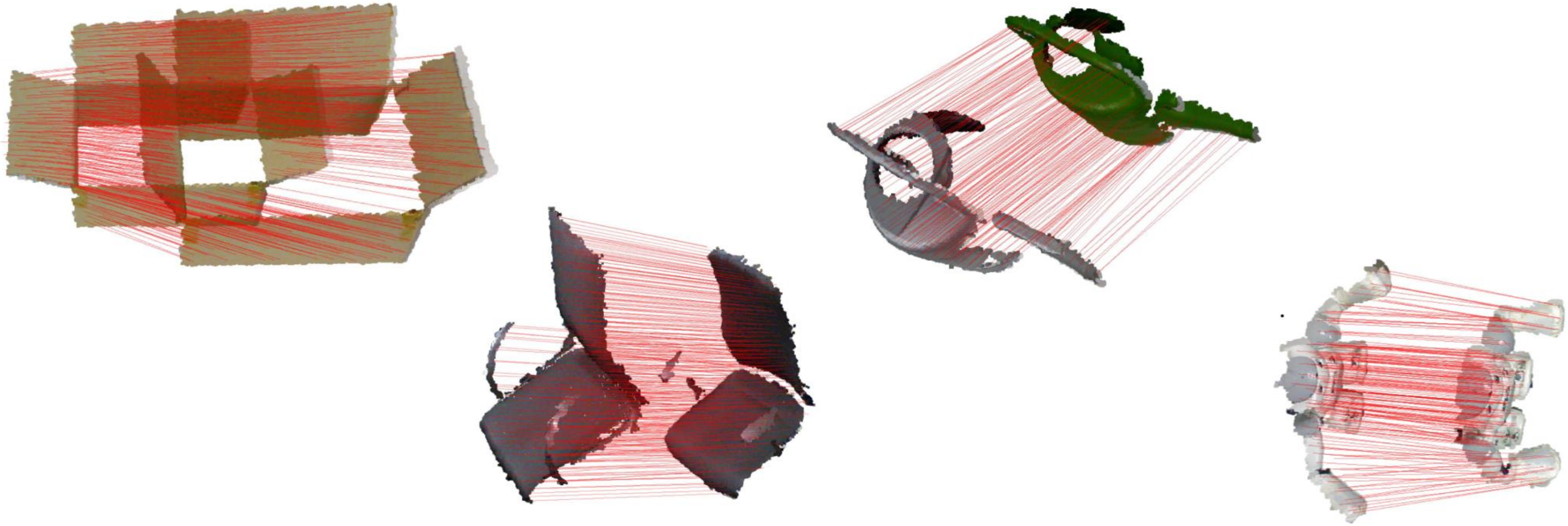


- Real-time registration with current RGB-D frame



Deformable RGB-D-Registration

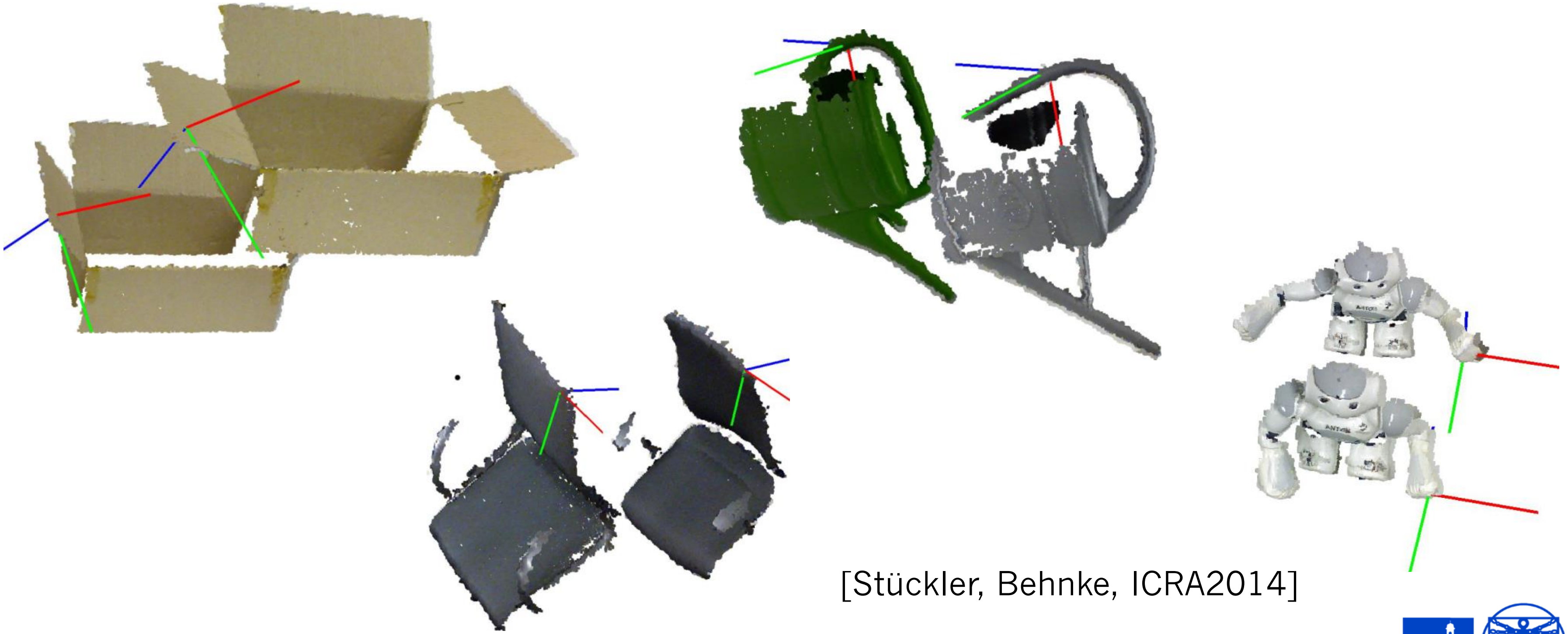
- Based on Coherent Point Drift method [Myronenko & Song, PAMI 2010]
- Multiresolution Surfel Map allows real-time registration



universitätbonn ais

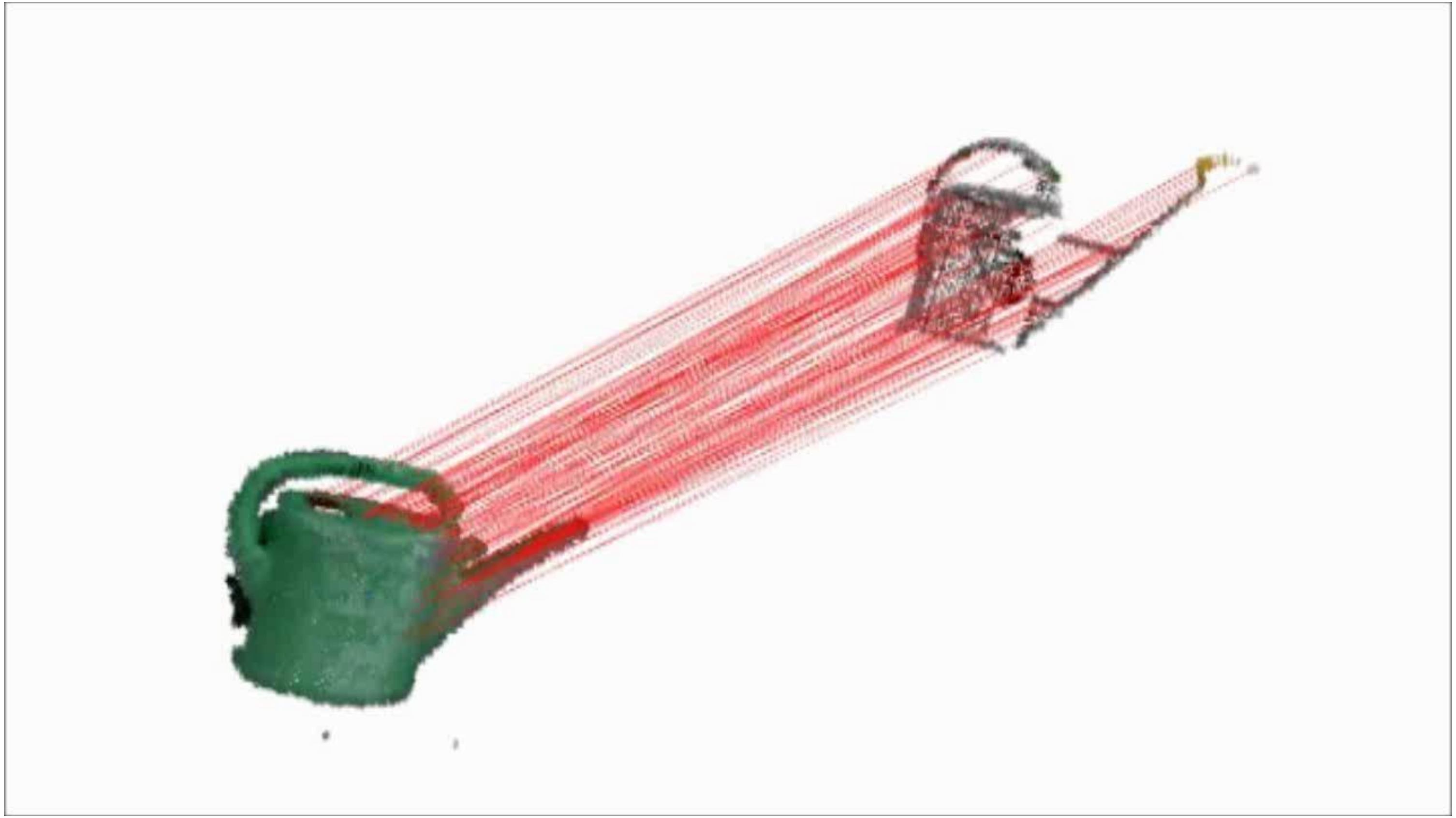
Transformation of Poses on Object

- Derived from the deformation field



[Stückler, Behnke, ICRA2014]

Grasp & Motion Skill Transfer

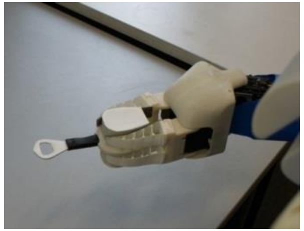


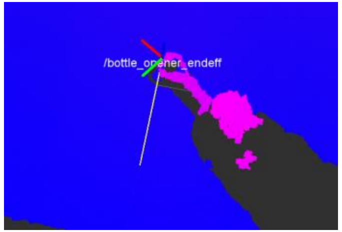
[Stückler, Behnke, ICRA2014]

universitätbonn ais

Tool use: Bottle Opener

- Tool tip perception





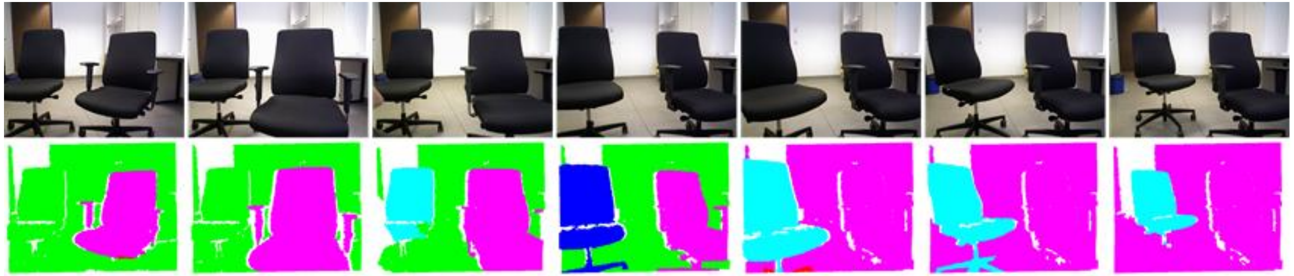
- Extension of arm kinematics
- Perception of crown cap
- Motion adaptation

[Stückler, Behnke, Humanoids 2014]

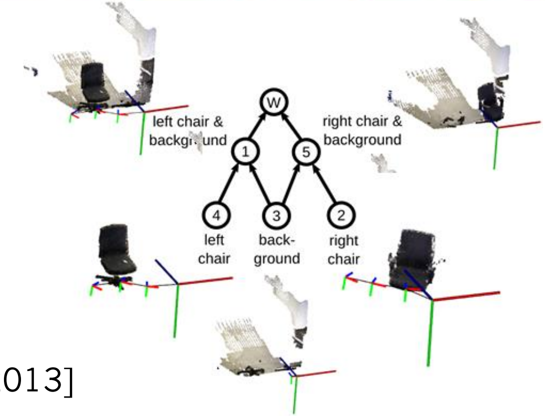
universitätbonn ais

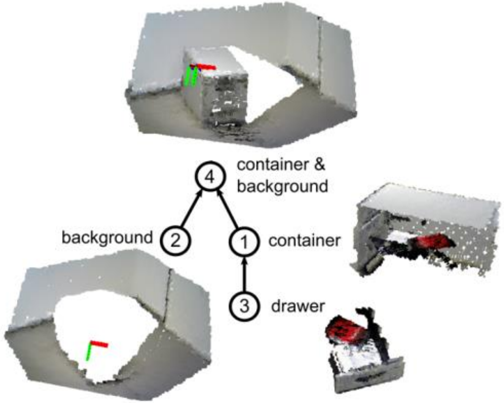
Hierarchical Object Discovery trough Motion Segmentation

- Simultaneous object modeling and motion segmentation



- Inference of a segment hierarchy



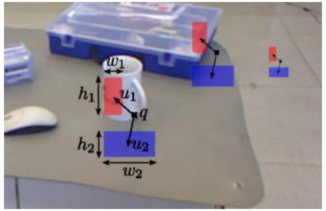


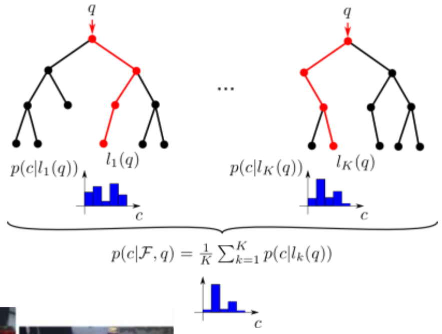
[Stückler, Behnke: IJCAI 2013]

universitätbonn ais

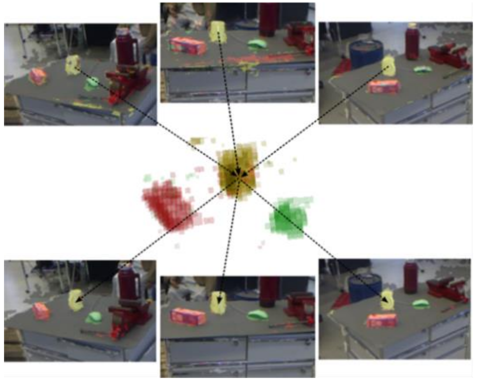
Semantic Mapping

- Pixel-wise classification of RGB-D images by random forests
- Compare color / depth of regions
- Size normalization
- 3D fusion through RGB-D SLAM
- Evaluation on NYU depth v2





[Stückler, Biresev, Behnke: IROS 2012]



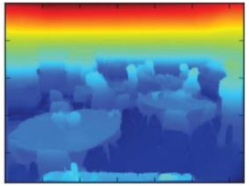
Ground truth

Segmentation

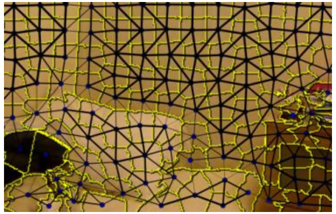
Accuracy in %	Ø Classes	Ø Pixels
Silberman et al. 2012	59,6	58,6
Couprie et al. 2013	63,5	64,5
Random forest	65,0	68,1
3D-Fusion	**66,8**	

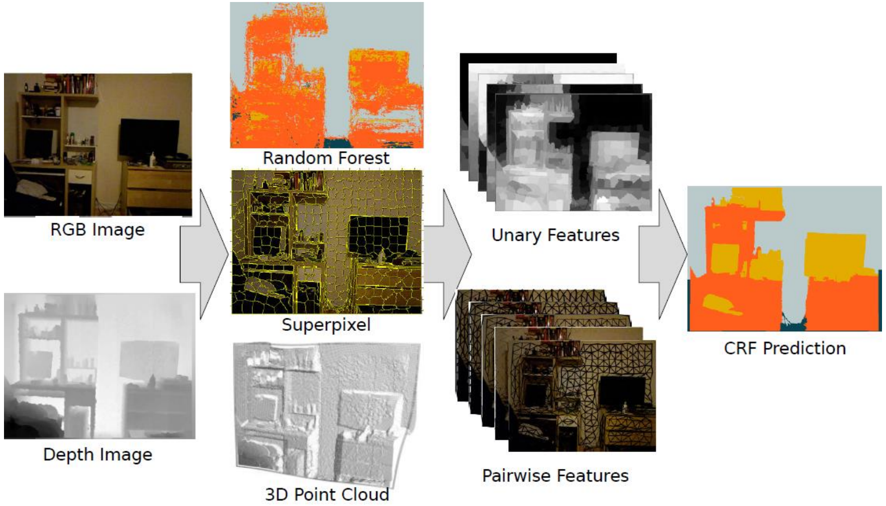
Learning Depth-sensitive CRFs

- SLIC+depth super pixels
- Unary features: random forest
- Height feature



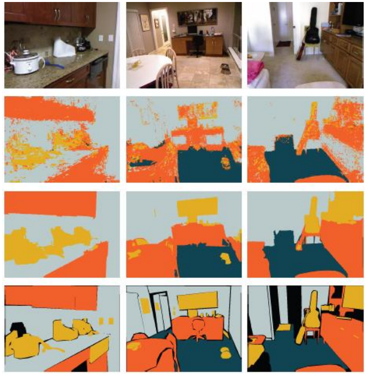
- Pairwise features
  - Color contrast
  - Vertical alignment
  - Depth difference
  - Normal differences





- Results:

	class average	pixel average
RF	65.0	68.3
RF + SP	65.7	70.1
RF + SP + SVM	70.4	70.3
RF + SP + CRF	**71.9**	**72.3**
Silberman *et al.*	59.6	58.6
Couprie *et al.*	63.5	64.5



Random forest

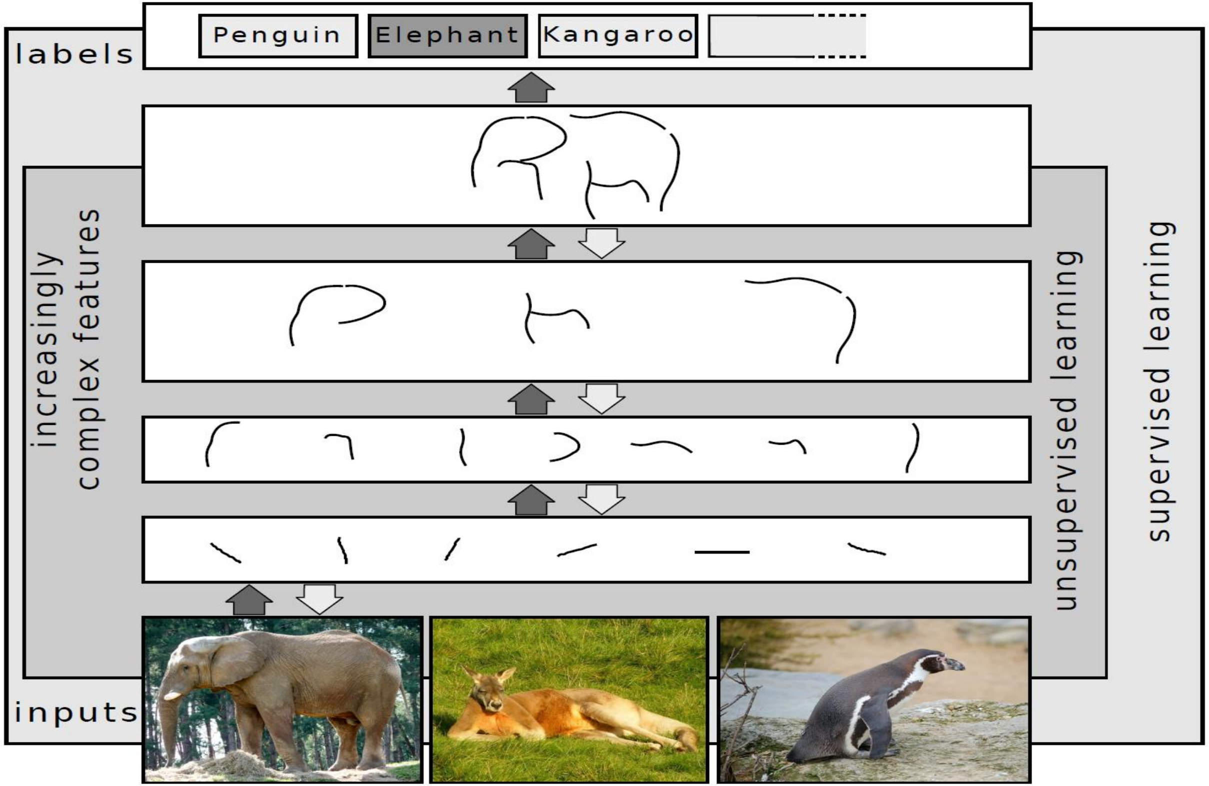
CRF prediction

Ground truth

universitätbonn ais

Deep Learning

- Learning layered represen-tations

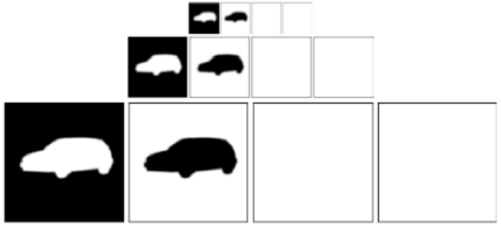


[Schulz; Behnke, KI 2012]

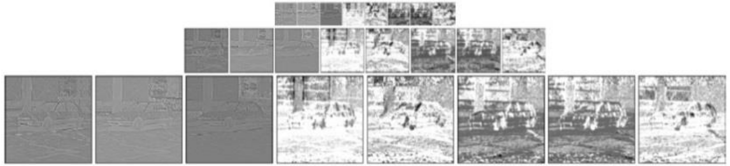
Object-class Segmentation

[Schulz, Behnke, ESANN 2012]

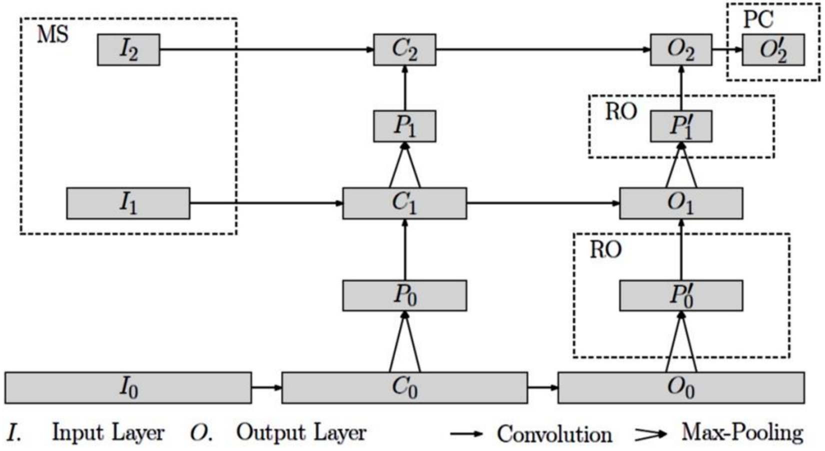
- Class annotation per pixel

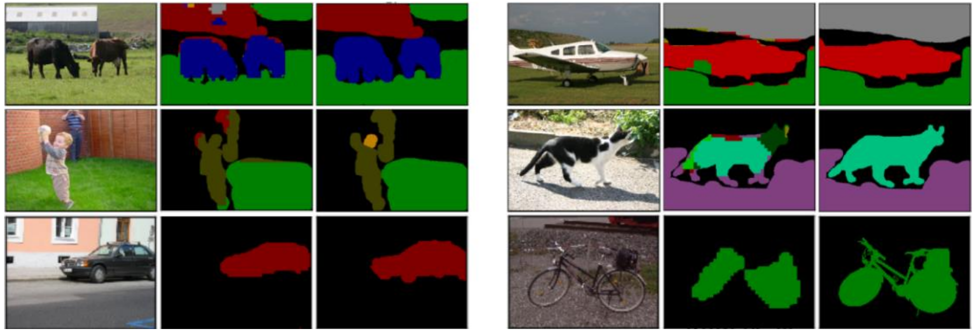


- Multi-scale input channels



- Evaluated on MSRC-9/21 and INRIA Graz-02 data sets





Input Output Truth

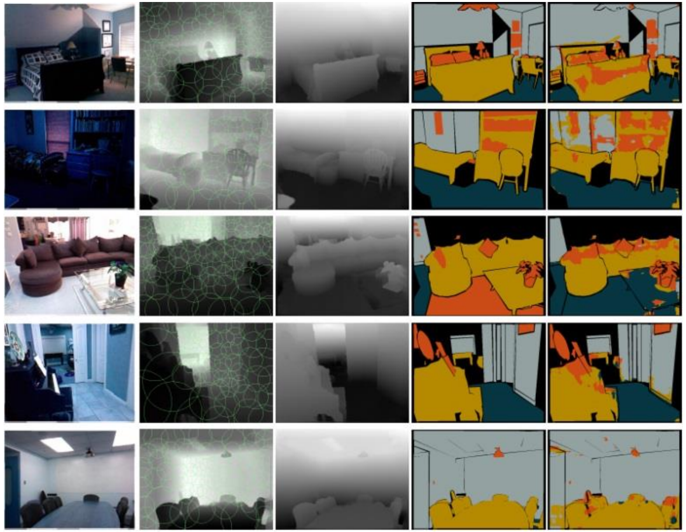
Object Detection in Natural Images

- Bounding box annotation
- Structured loss that directly maximizes overlap of the prediction with ground truth bounding boxes
- Evaluated on two of the Pascal VOC 2007 classes

[Schulz, Behnke, ICANN 2014]

RGB-D Object-Class Segmentation

- Covering windows segmented with CNN
- Scale input according to depth, compute pixel hight



RGB Depth Height Truth Output

Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	94.9	78.9	71.1	42.7	71.9	72.3
Random Forest [8]	90.8	81.6	67.9	19.9	65.1	68.3
Couprie et al.[9]	87.3	86.1	45.3	35.5	63.6	64.5
Höft et al.[10]	77.9	65.4	55.9	49.9	62.3	62.0
Silberman [12]	68	59	70	42	59.7	58.6

CW is covering windows, H is height above ground, DN is depth normalized patch sizes. SP is averaged within superpixels and SVM-reweighted. CRF is a conditional random field over superpixels [8]. Structure class numbers are optimized for class accuracy.

[Schulz, Höft, Behnke, ESANN 2015]

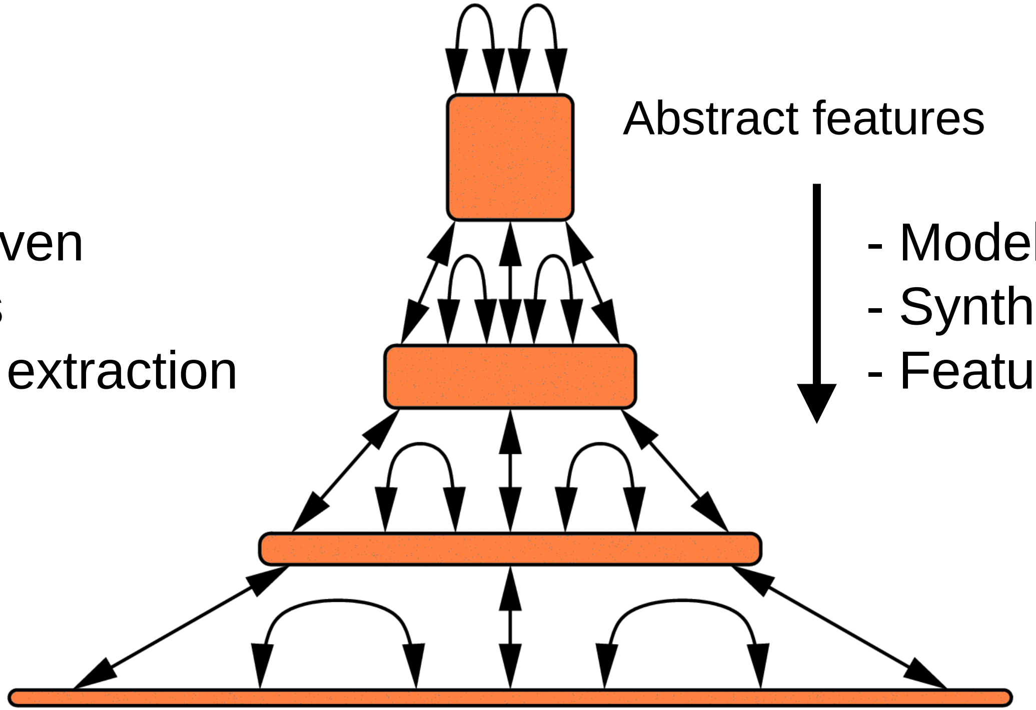
universitätbonn ais

Neural Abstraction Pyramid

[Behnke, Rojas, IJCNN 1998]
[Behnke, LNCS 2766, 2003]

Abstract features

- Data-driven
- Analysis
- Feature extraction



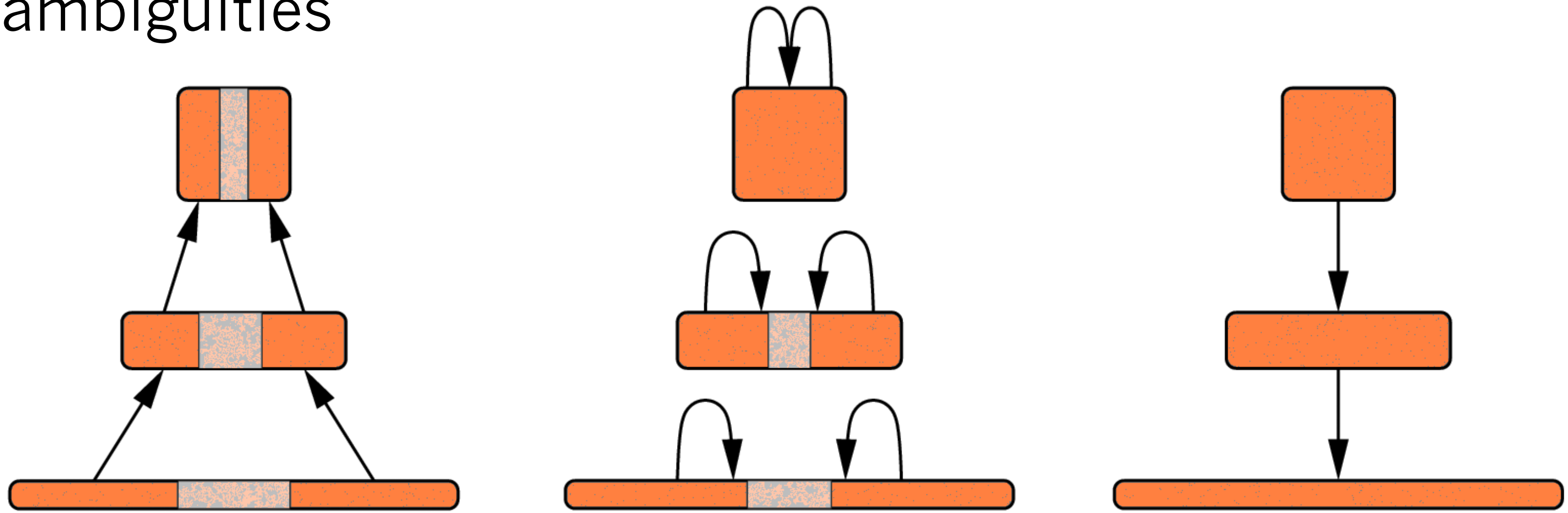
- Model-driven
- Synthesis
- Feature expansion

Signals

- Grouping - Competition - Completion

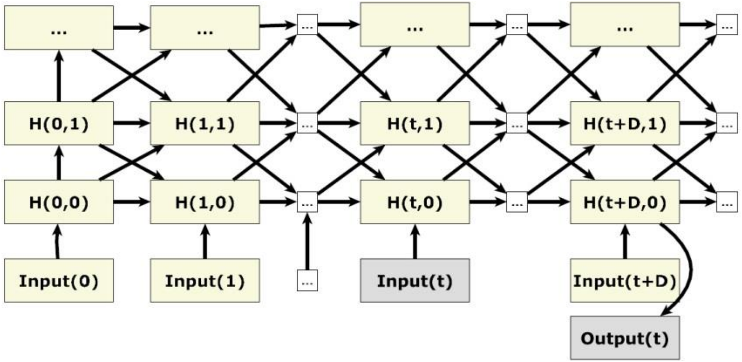
Iterative Image Interpretation

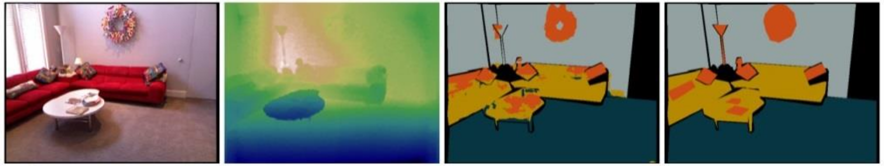
- Interpret most obvious parts first
- Use partial interpretation as context to resolve local ambiguities



Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- NYU Depth V2 contains RGB-D video sequences
- Recursive computation is efficient for temporal integration





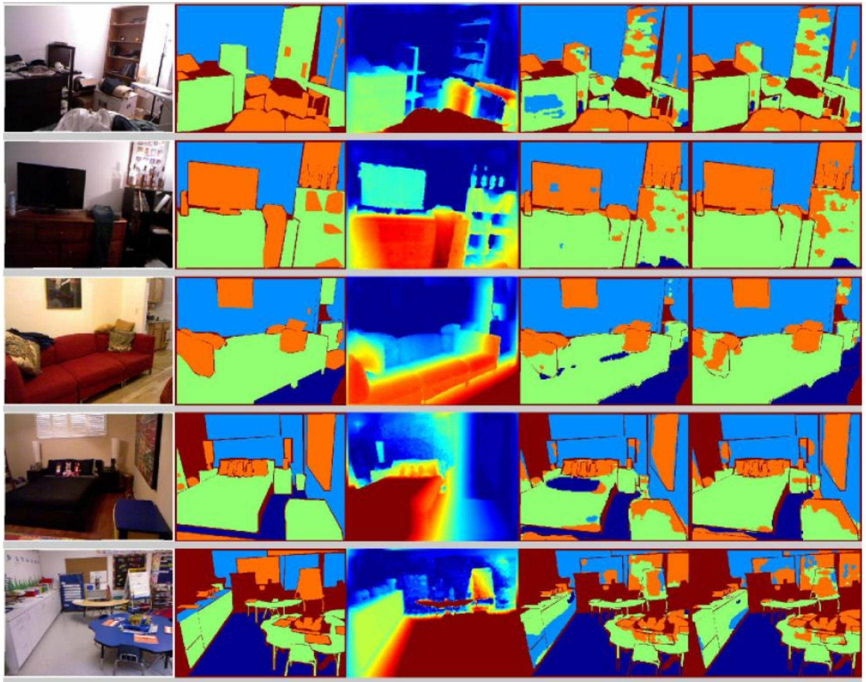
RGB Depth Output Truth

Method	Class Accuracies (%) ground	struct	furnit	prop	Average (%) Class	Pixel
Höft *et al.* [19]	77.9	65.4	55.9	49.9	62.0	61.1
Unidirectional + MS	73.4	66.8	**60.3**	49.2	62.4	63.1
Schulz *et al.* [20] (no height)	87.7	70.8	57.0	53.6	67.3	65.5
Unidirectional + SW	**90.0**	**76.3**	52.1	**61.2**	**69.9**	**67.5**

[Pavel, Schulz, Behnke, IJCNN 2015]

Geometric and Semantic Features for RGB-D Object-class Segmentation

- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly

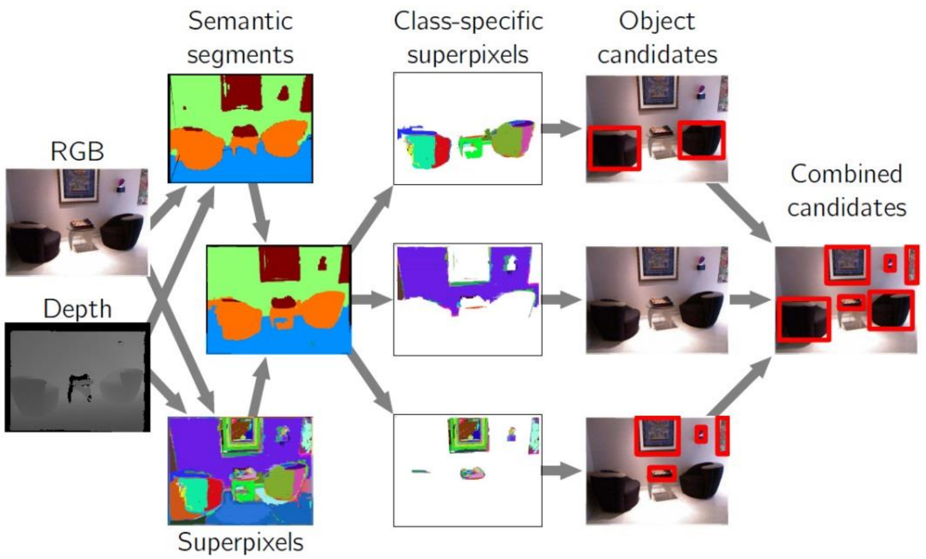


RGB Truth DistWall OutWO OutWithDistWall

[Husain et al. RA-L 2016]

Semantic Segmentation Priors for Object Discovery

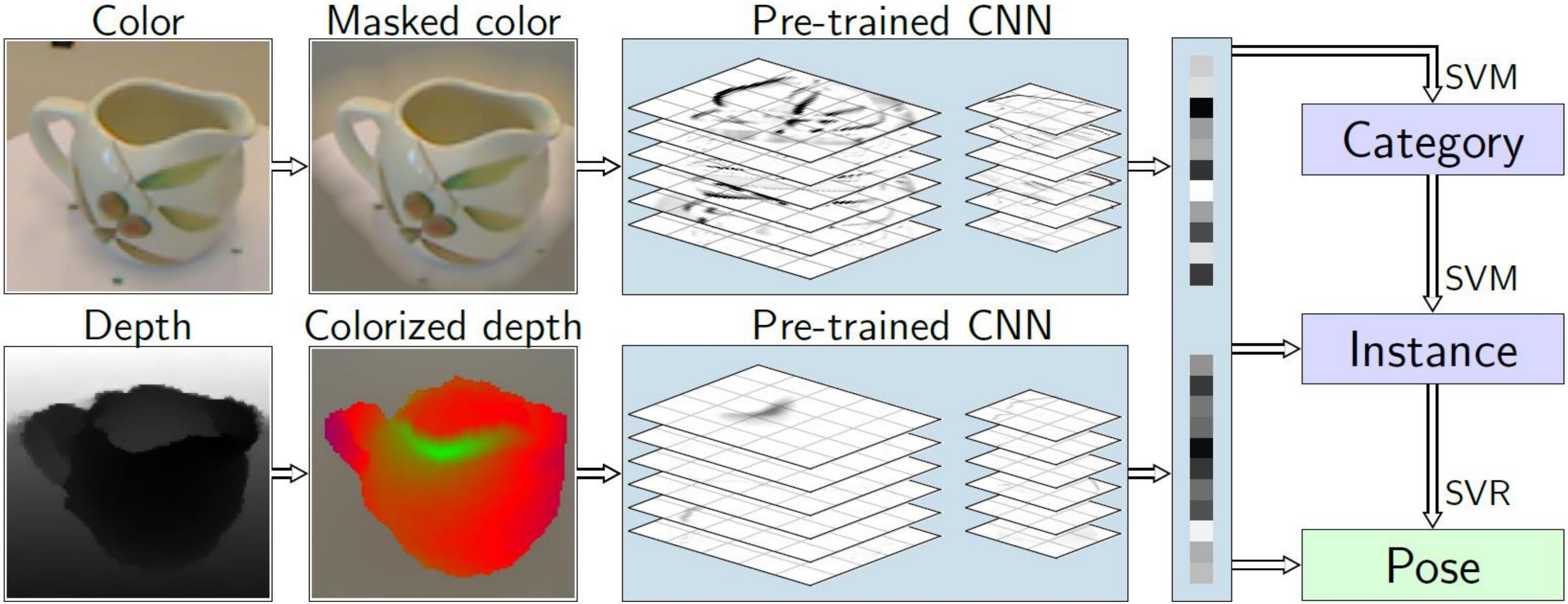
- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders



[Garcia et al. under review]

universitätbonn ais

RGB-D Object Recognition and Pose Estimation

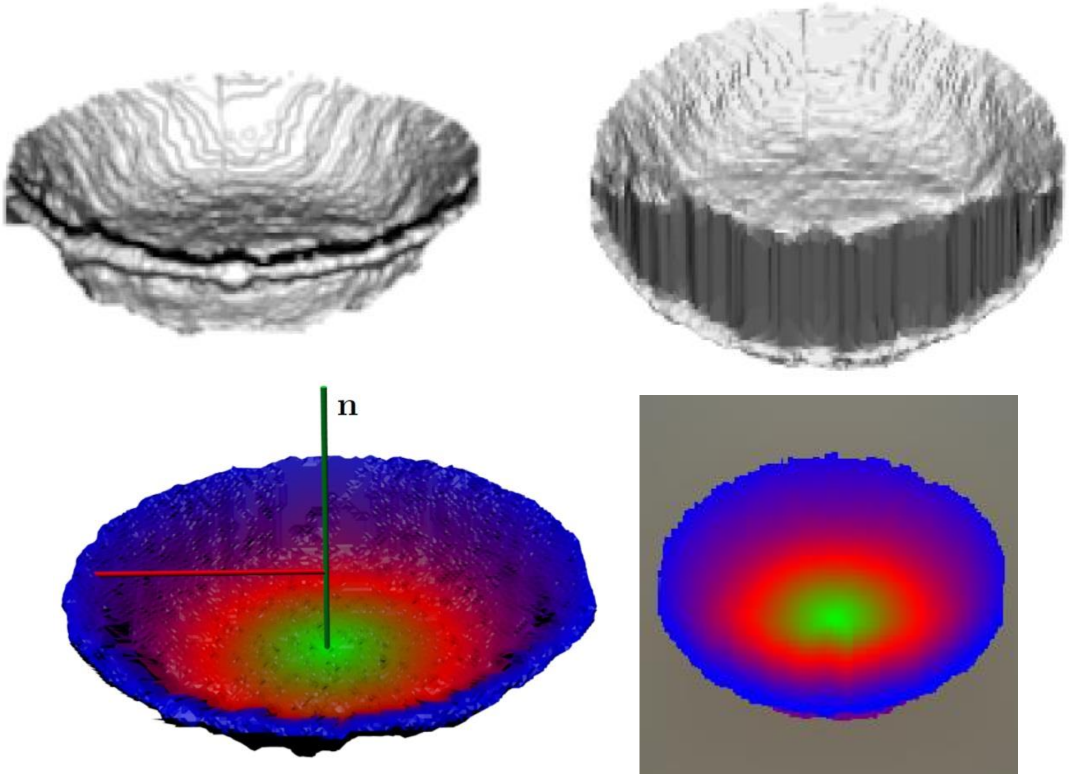


[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view

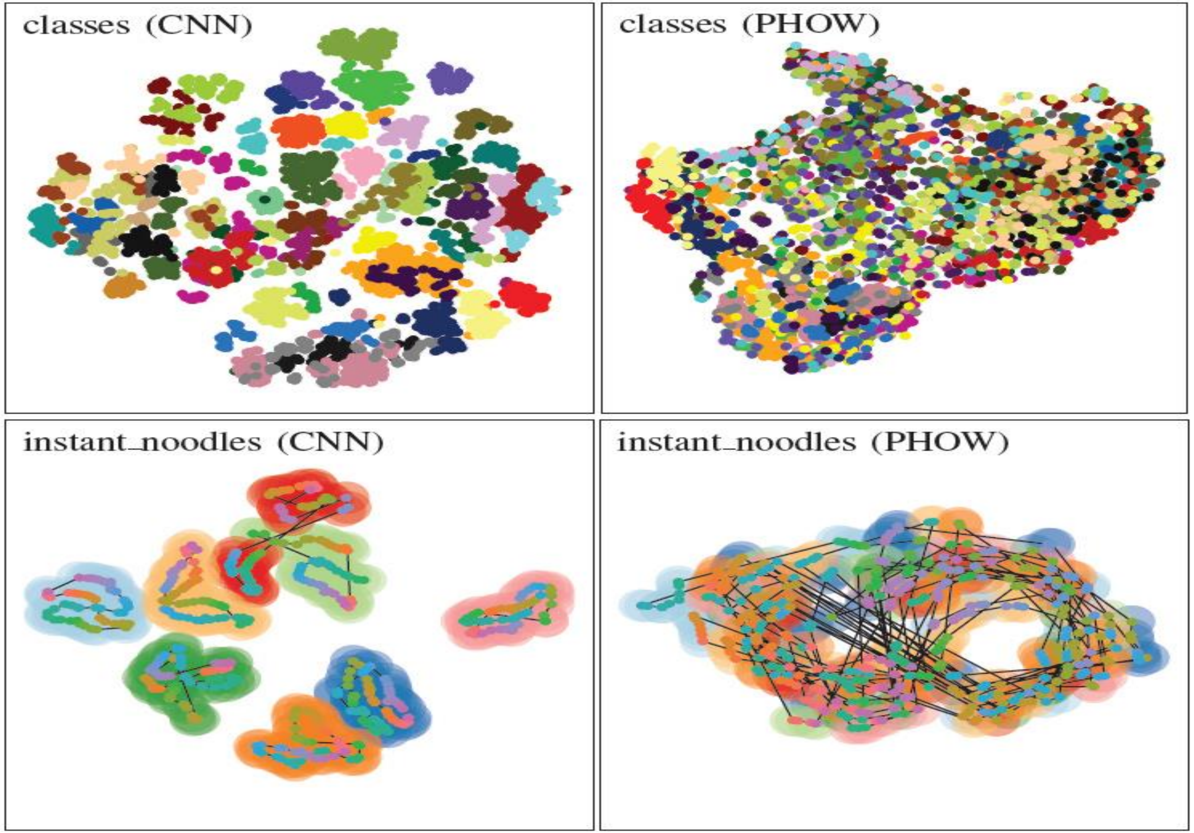
- Colorization based on distance from center vertical



[Schwarz, Schulz, Behnke, ICRA2015]

Pretrained Features Disentangle Data

- t-SNE embedding



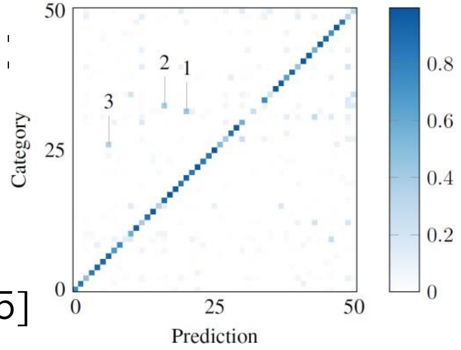
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

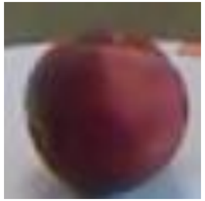
Method	Category Accuracy (%) RGB	Category Accuracy (%) RGB-D	Instance Accuracy (%) RGB	Instance Accuracy (%) RGB-D
Lai *et al.* [1]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9
Bo *et al.* [2]	82.4 ± 3.1	87.5 ± 2.9	**92.1**	92.8
PHOW[3]	80.2 ± 1.8	—	62.8	—
Ours	$\mathbf{83.1 \pm 2.0}$	88.3 ± 1.5	92.0	**94.1**
Ours	$\mathbf{83.1 \pm 2.0}$	$\mathbf{89.4 \pm 1.3}$	92.0	**94.1**

- Confusion



1: pitcher / coffe mug

2: peach / sponge



[Schwarz, Schulz, Behnke, ICRA2015]

universitätbonn ais

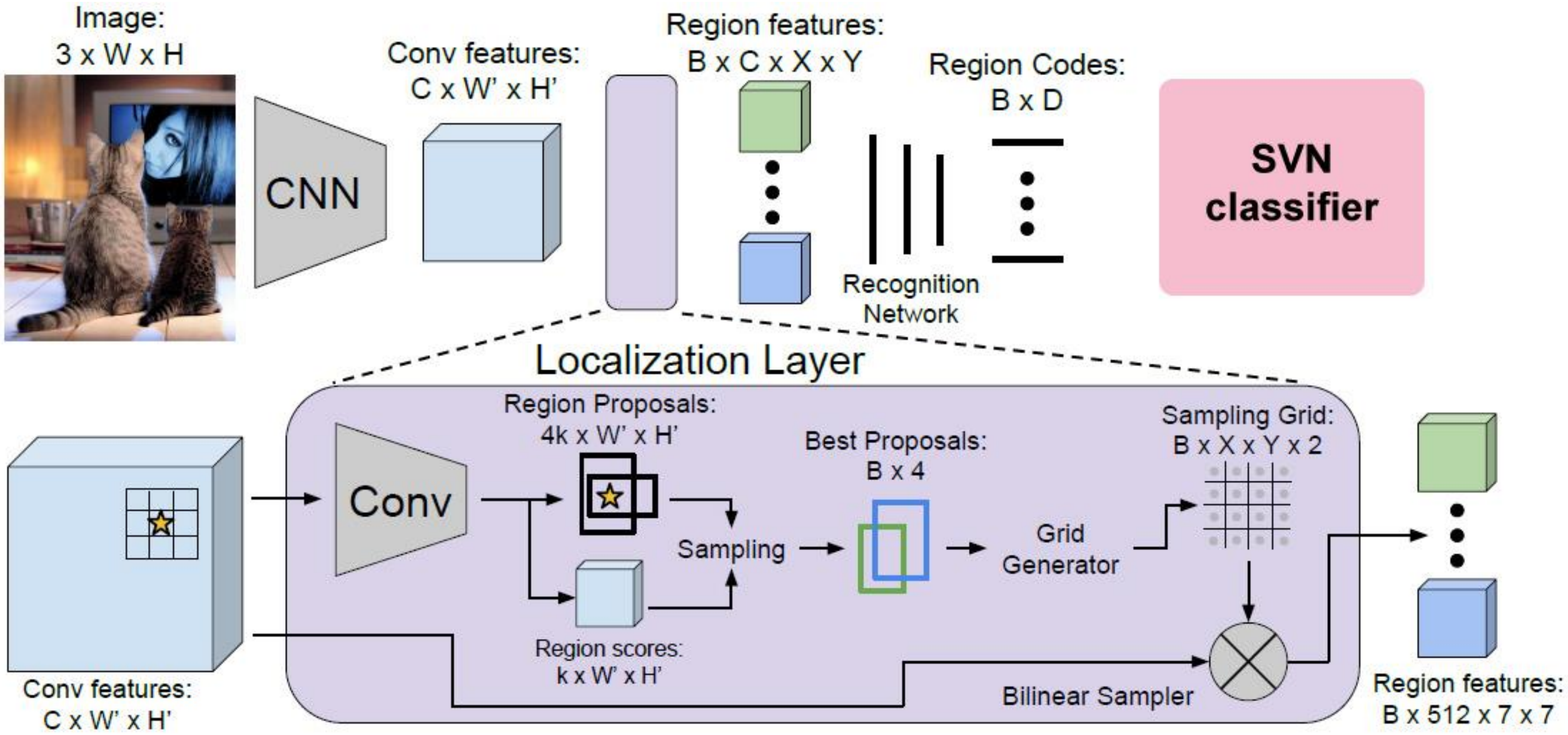
Amazon Picking Challenge

- Large variety of objects
- Unordered in shelf or tote
- Picking and stowing tasks

Deep Learning Semantic Segmentation

- Adapted from our segmentation of indoor scenes [Husain et al. RA-L 2016]

DenseCap Object Detection + SVM

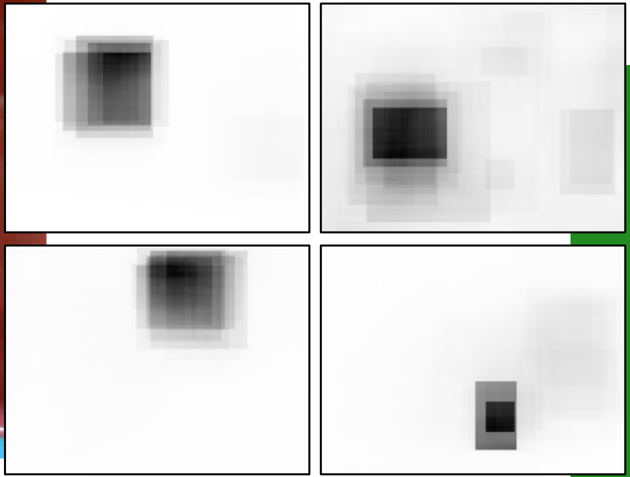


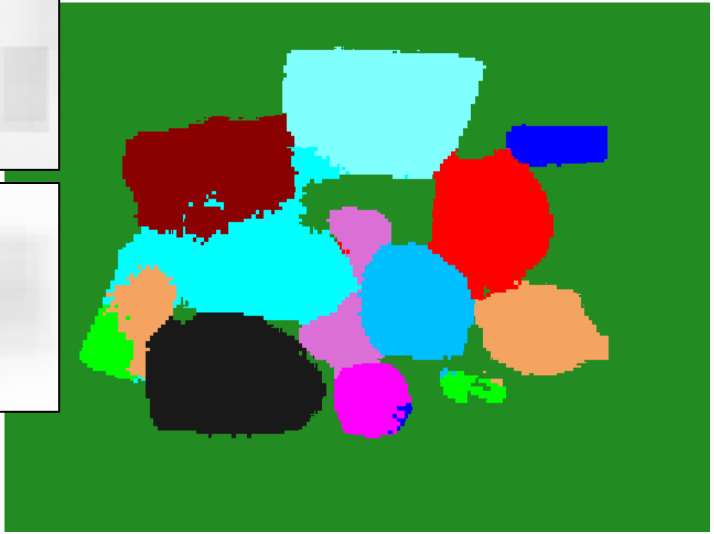
[Johnson et a CVPR 2016]

Combined with Object Detections

- DenseCap features and SVM classifier

[Johnson et a CVPR 2016]





Challenge Will Start Tomorrow

Post Scriptum: NimbRo Picking Results

- 2^{nd} Place Stowing (186 points)
- 3^{rd} Place Picking (97 points)

Conclusions

- Semantic perception is challenging
- Simple methods rely on strong assumptions
- Depth helps with segmentation, allows for size normalization, geometric features, shape descriptors
- Deep learning methods work well
- Transfer of features from large data sets
- Many open problems, e.g. total scene understanding, incorporating physics, ...

Questions?