Semantic RGB-D Perception for Cognitive Robots

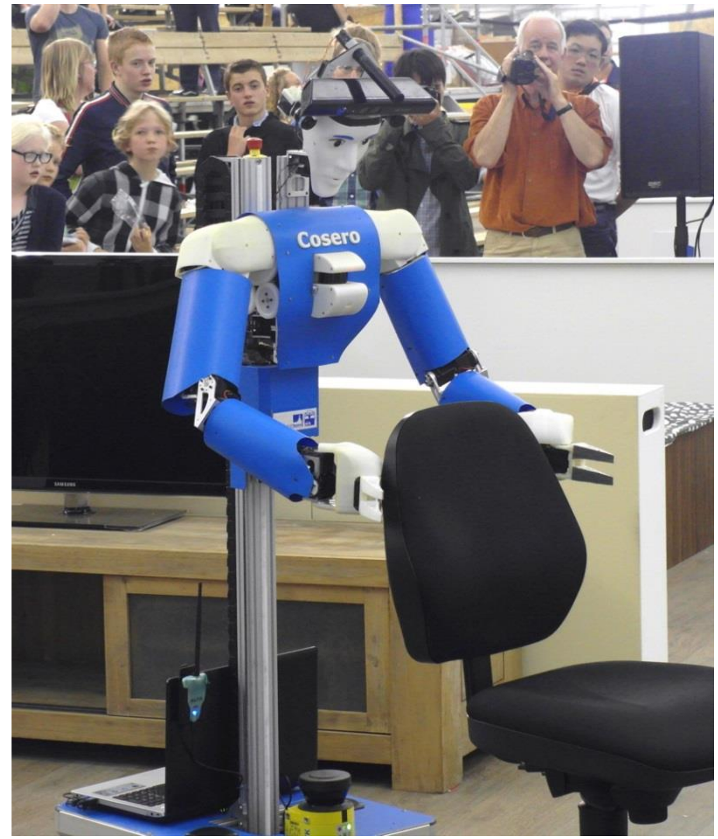
Sven Behnke

Computer Science Institute VI
Autonomous Intelligent Systems

universitätbonn ais

Our Domestic Service Robots

Dynamaid



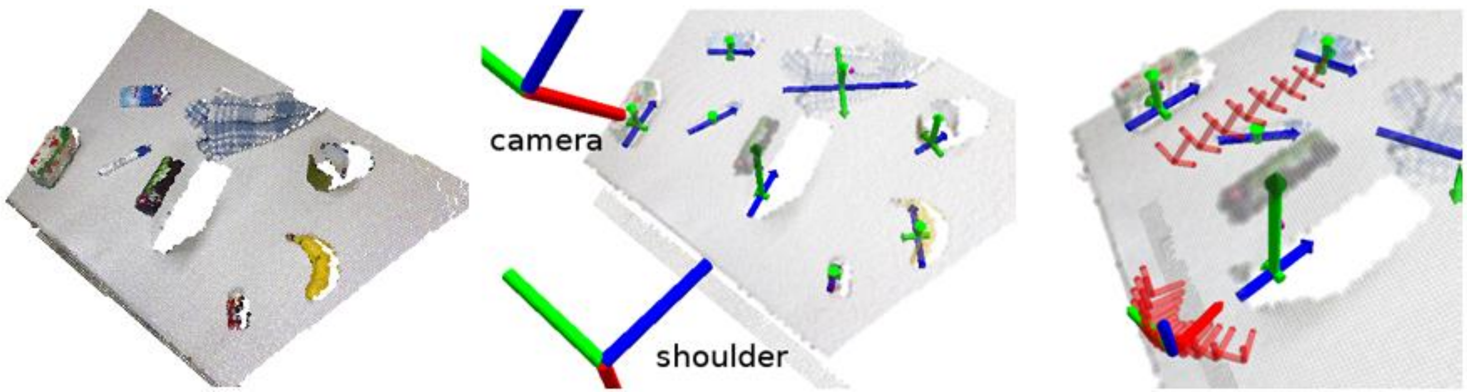
Cosero

- Size: 100-180 cm, weight: 30-35 kg
- 36 articulated joints
- PC, laser scanners, **Kinect**, microphone, ...

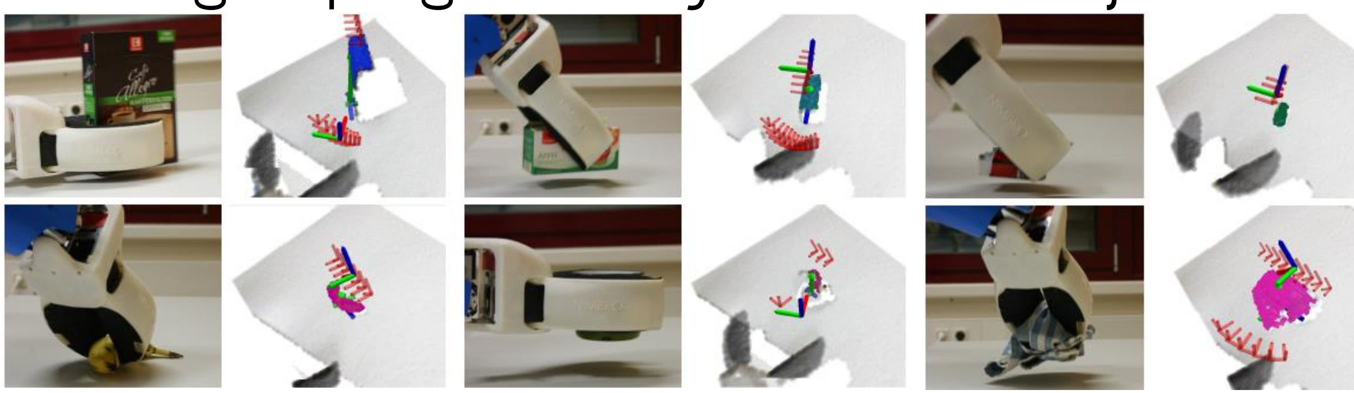
RoboCup 2013 Eindhoven

Analysis of Table-top Scenes and Grasp Planning

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

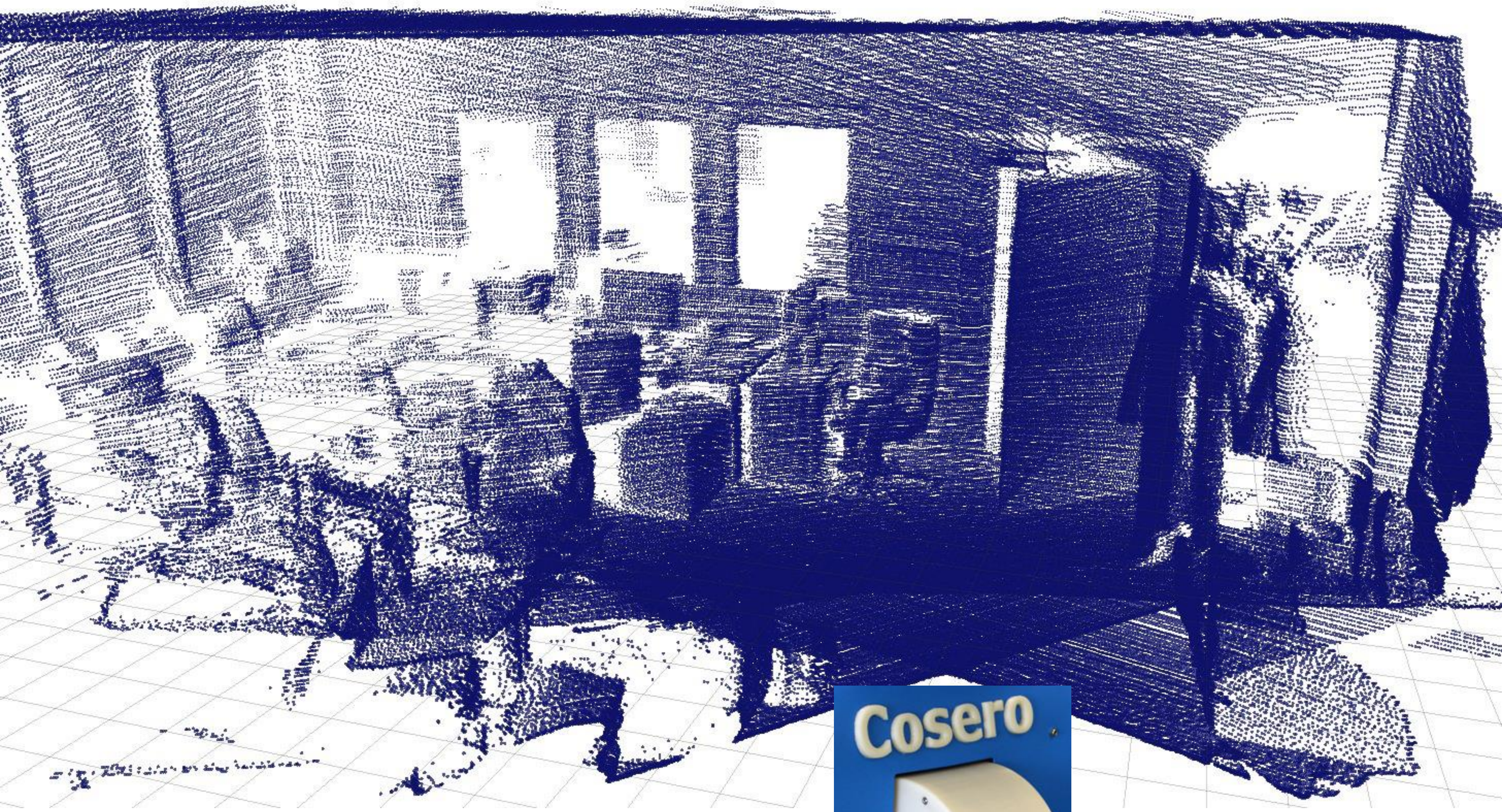


- Flexible grasping of many unknown objects

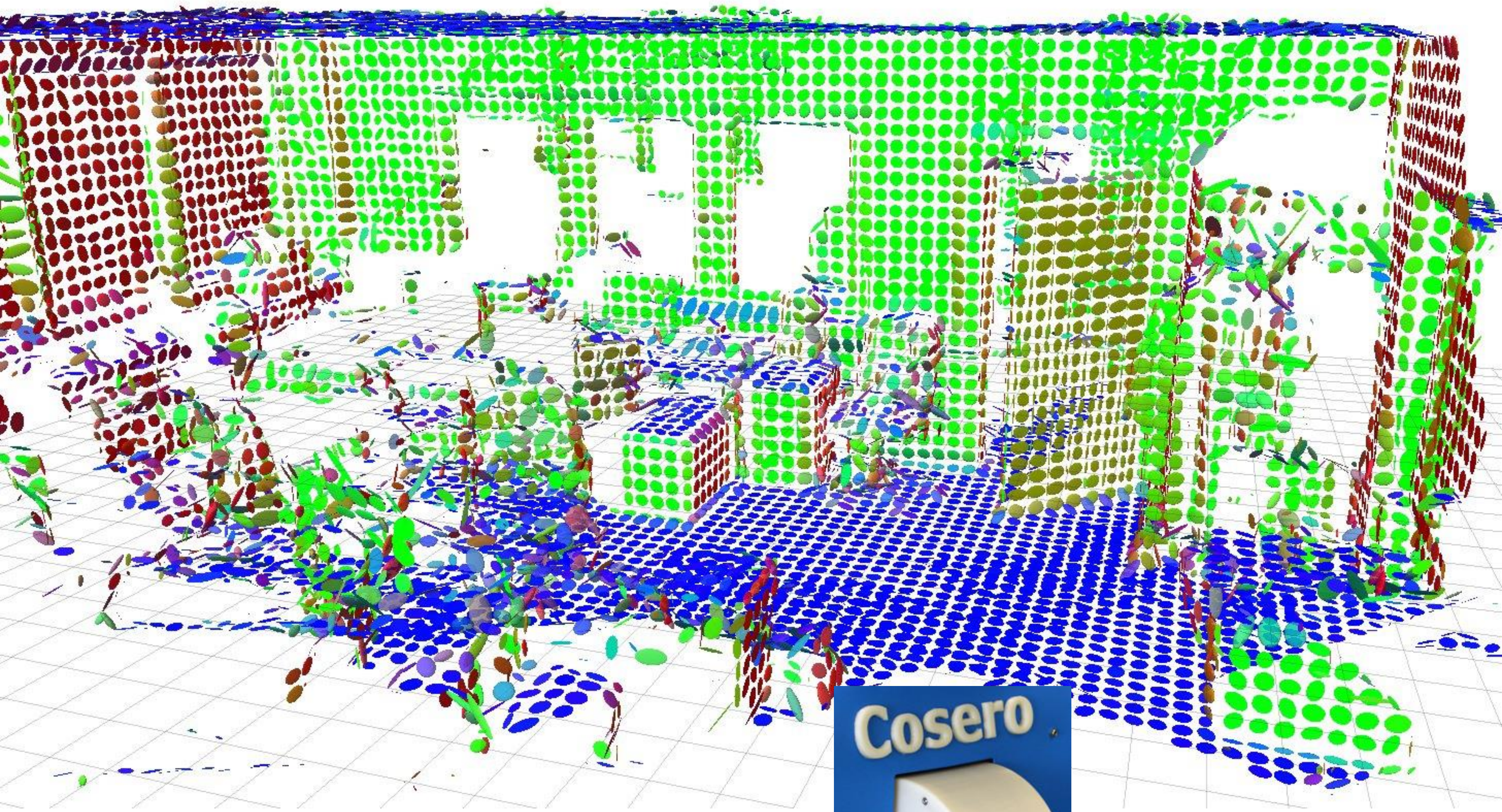


[Stückler, Steffens, Holz, Behnke, Robotics and Autonomous Systems 2012]

3D-Mapping with Surfels

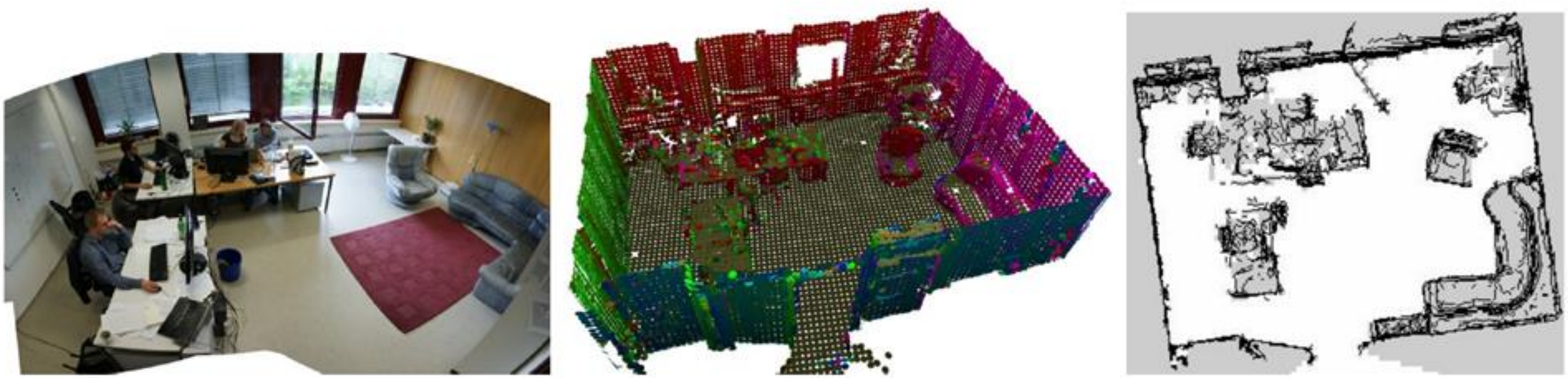


3D-Mapping with Surfels



3D-Mapping and Localization

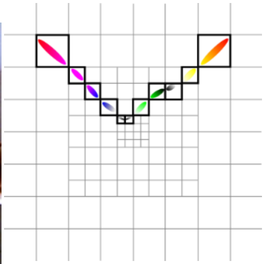
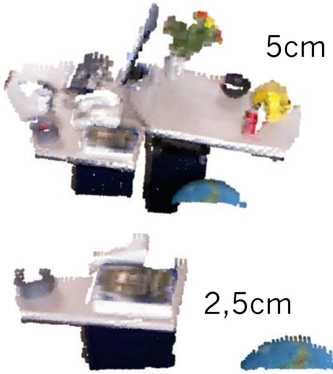
- Registration of 3D laser scans
- Representation of point distributions in voxels
- Drivability assessment trough region growing
- Robust localization using 2D laser scans



[Kläß, Stückler, Behnke: Robotik 2012]

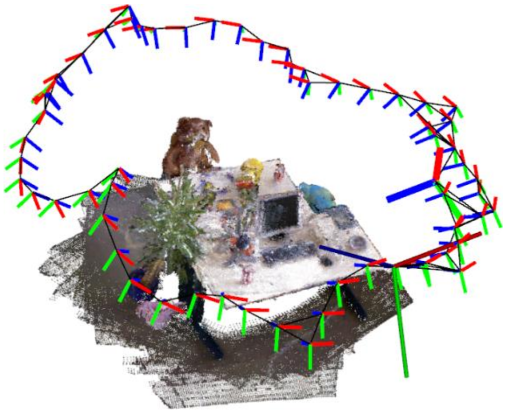
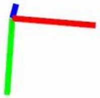
3D Mapping by RGB-D SLAM

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

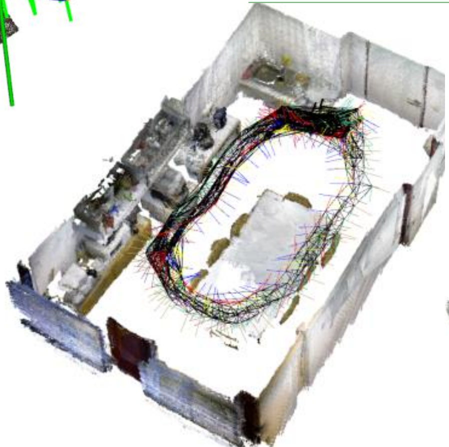
[Stückler, Behnke:
Journal of Visual Communication
and Image Representation 2013]

- Global optimization

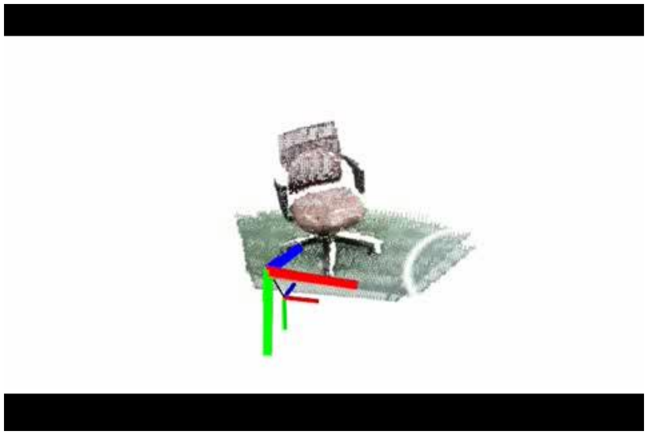
- Multi-camera SLAM

[Stoucken, Diplomarbeit 2013]

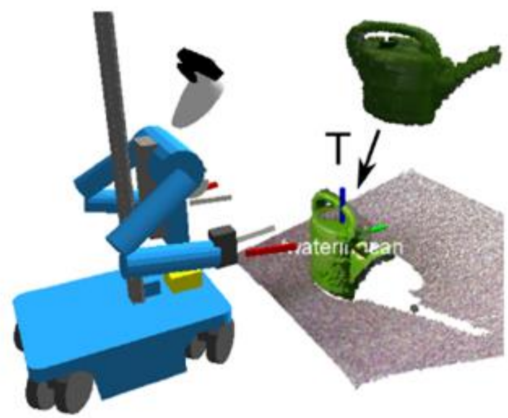



Learning and Tracking Object Models

- Modeling of objects by RGB-D-SLAM

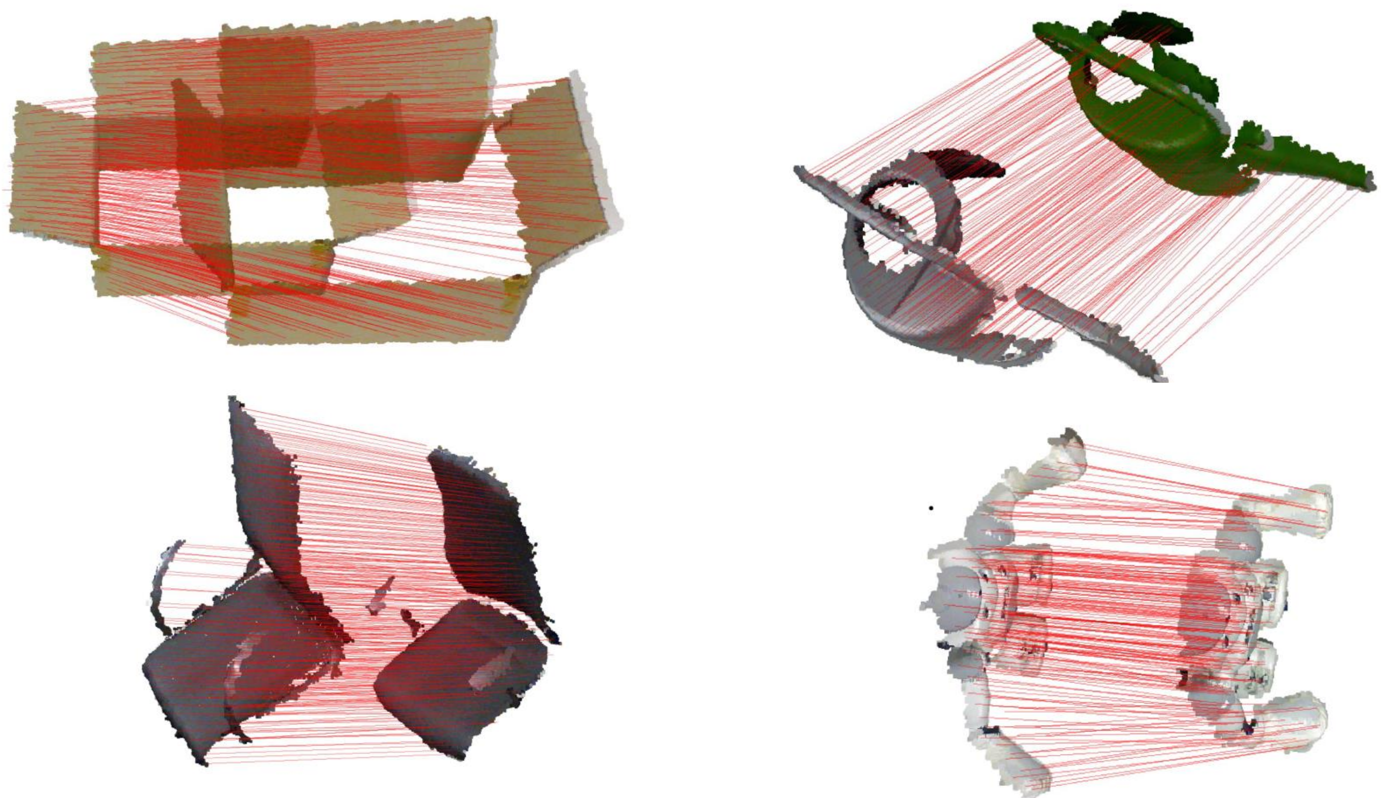



- Real-time registration with current RGB-D image

Deformable RGB-D-Registration

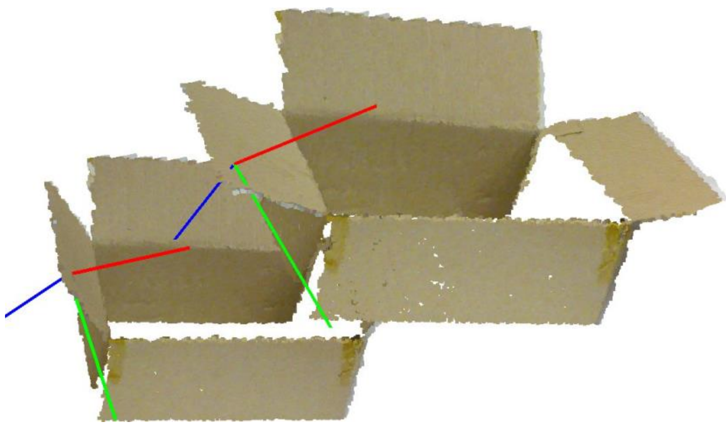
- Based on Coherent Point Drift method [Myronenko & Song, PAMI 2010]
- Multiresolution Surfel Map allows real-time registration

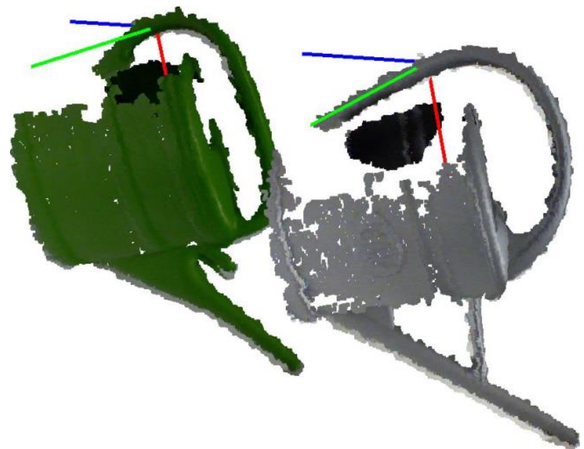


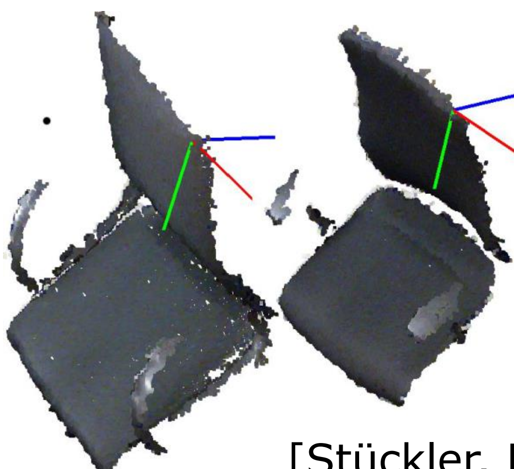
[Stückler, Behnke, ICRA2014]

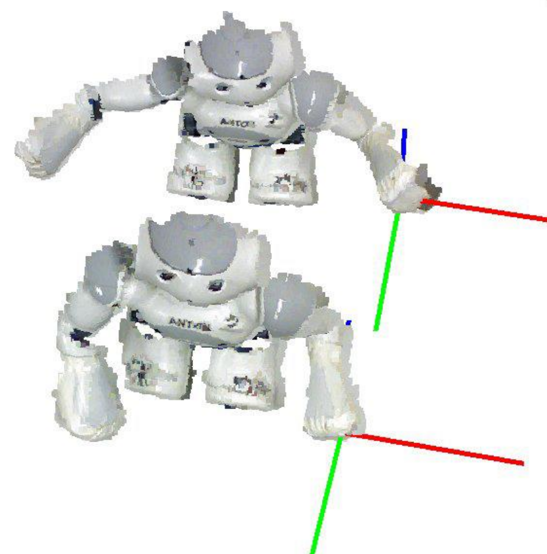
Transformation of Poses on Object

- Derived from the deformation field



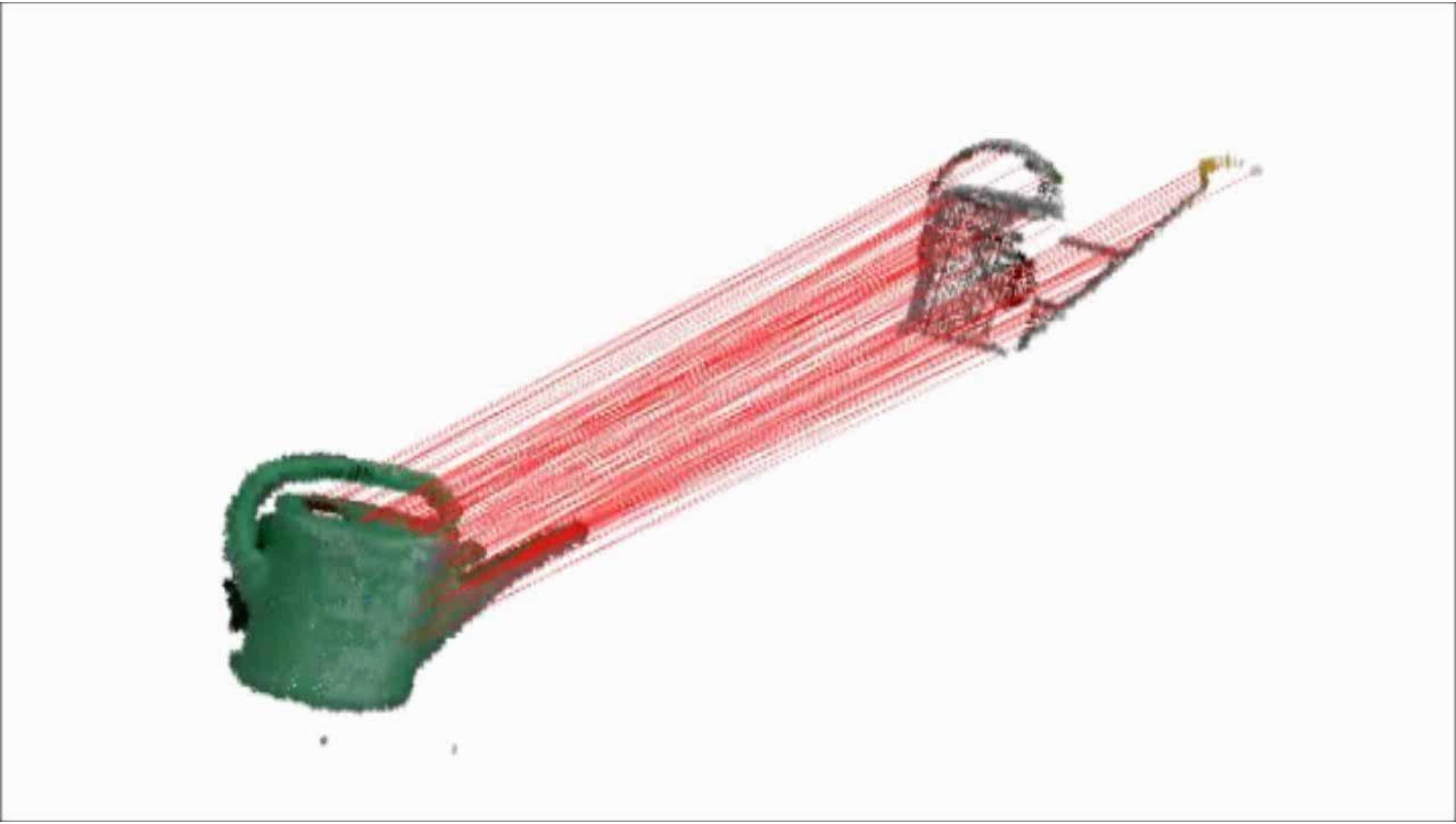






[Stückler, Behnke, ICRA2014]

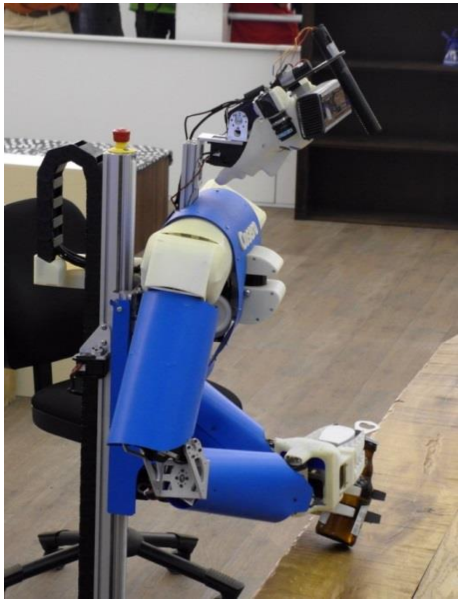
Grasp & Motion Skill Transfer

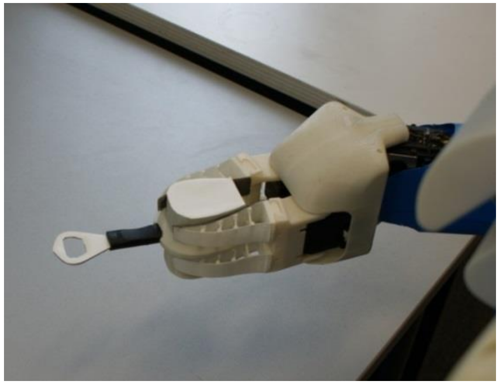


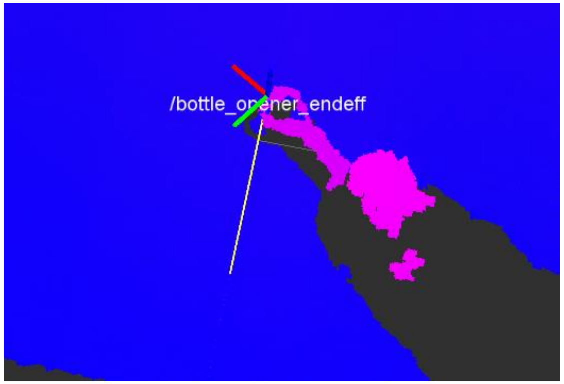
- Demonstration at RoboCup 2013 [Stückler, Behnke, ICRA2014]

Tool use: Bottle Opener

- Tool tip perception
- Extension of arm kinematics
- Perception of crown cap
- Motion adaptation







[Stückler, Behnke, Humanoids 2014]

Picking Sausage, Bimanual Transport

- Perception of tool tip and sausage
- Alignment with main axis of sausage

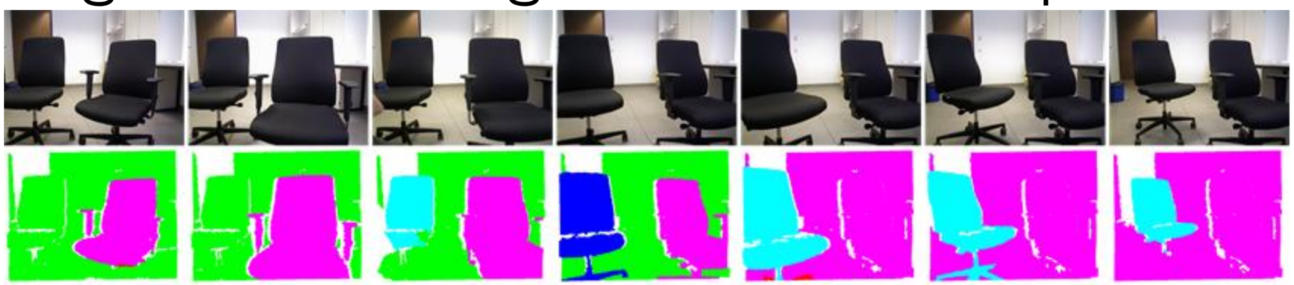
- Our team NimbRo won the RoboCup@Home League in three consecutive years

[Stückler, Behnke, Humanoids 2014]

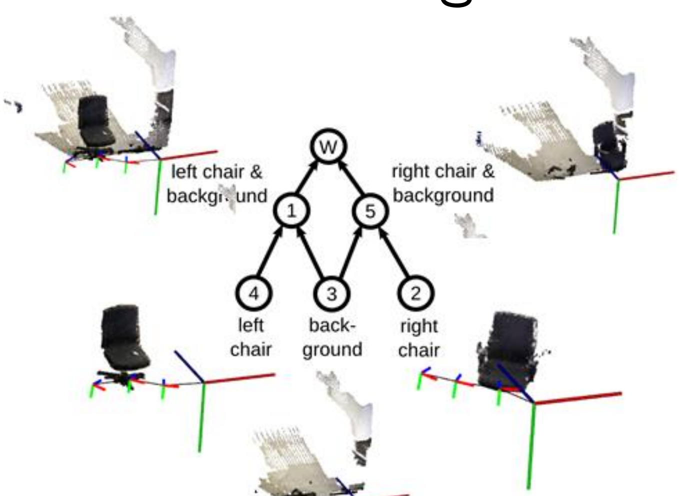
Hierarchical Object Discovery trough Motion Segmentation

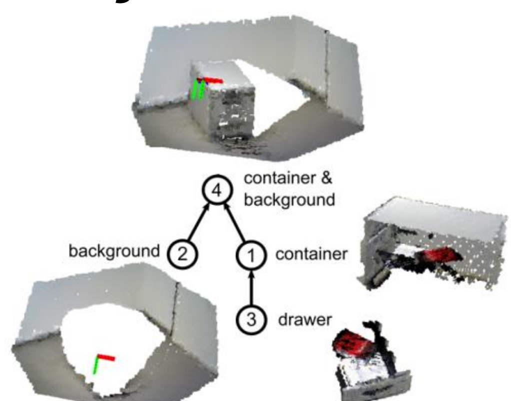
- Motion is strong segmentation cue
- Both camera and object motion

- Segment-wise registration of a sequence



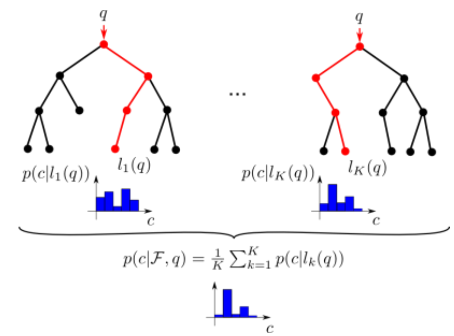
- Inference of a segment hierarchy



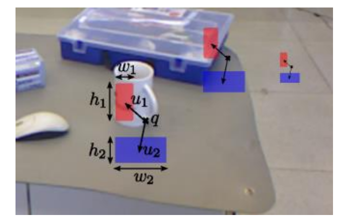


[Stückler, Behnke: IJCAI 2013]

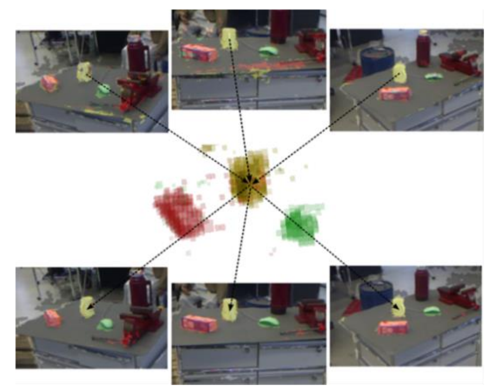
Semantic Mapping



- Pixel-wise classification of RGB-D images by random forests
- Inner nodes compare color / depth of regions
- Size normalization
- Training and recall on GPU
- 3D fusion through RGB-D SLAM
- Evaluation on own data set and NYU depth v2



[Stückler, Biresev, Behnke: IROS 2012]



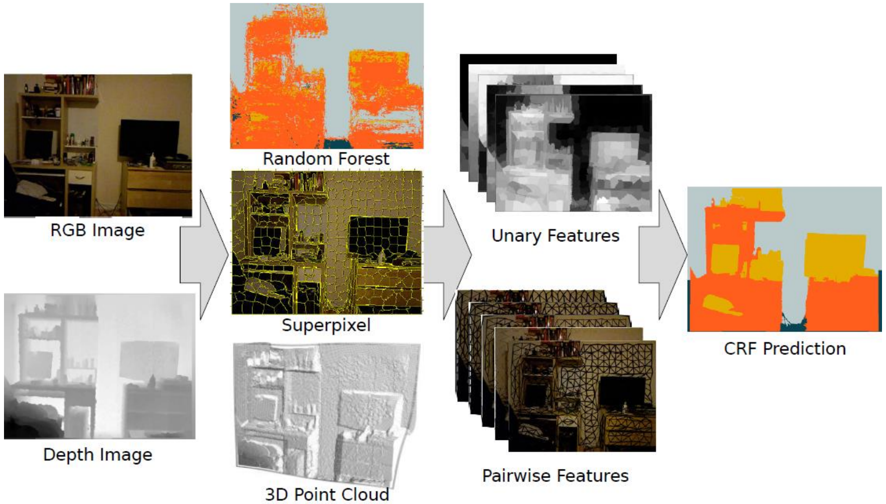
Ground truth

Segmentation

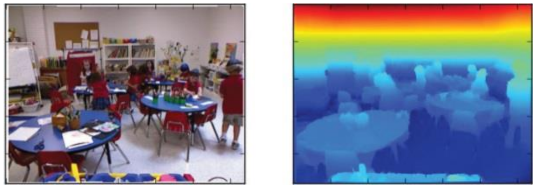
[Stückler et al., Journal of Real-Time Image Processing 2014]

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**	**70,6**

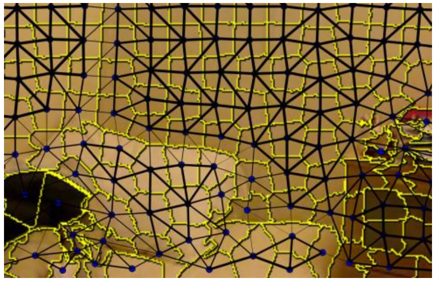
Learning Depth-sensitive CRFs



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



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



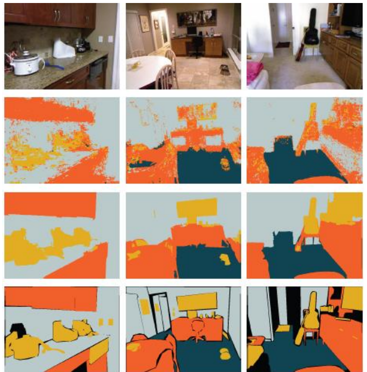
similarity between superpixel normals

- Results:

Random forest

CRF prediction

Ground truth

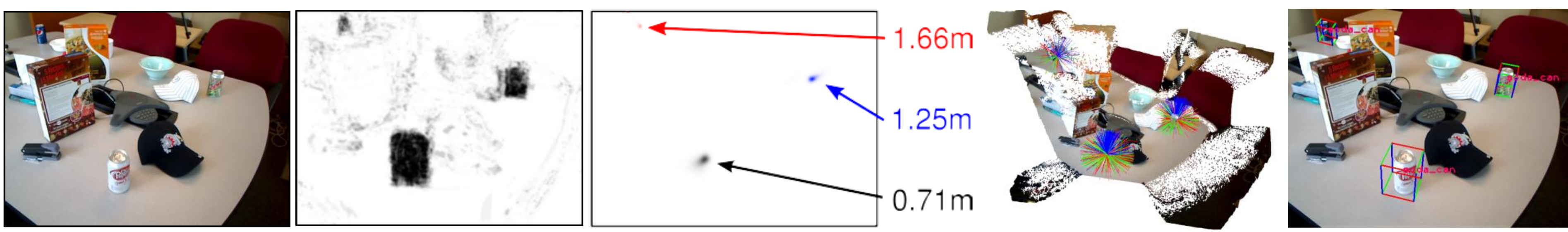


	class average	pixel average
RF	65.0	68.3
RF + SP	65.7	70.1
RF + SP + SVM	70.4	70.3
RF + SP + CRF	**71.9**	**72.3**
Silberman *et al.*	59.6	58.6
Couprie *et al.*	63.5	64.5

[Müller and Behnke, ICRA 2014]

Object Class Detection in RGB-D

- Hough forests make not only object class decision, but describe object center
- RGB-D objects data set
- Color and depth features
- Training with rendered scenes
- Detection of object position and orientation



Scene | Class prob. | Object centers | Orientation | Detected objects

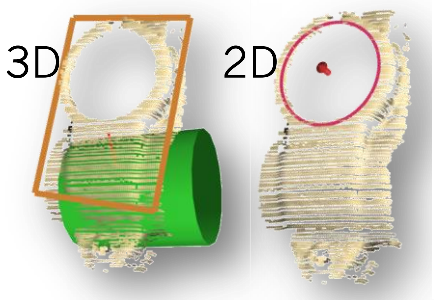
- Depth helps a lot

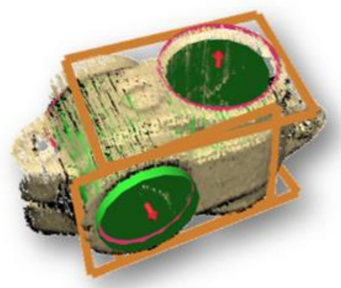
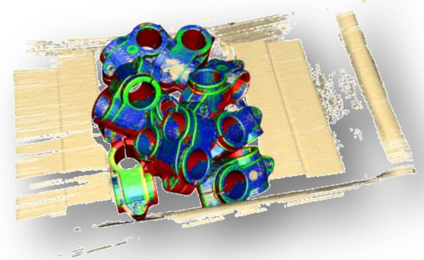
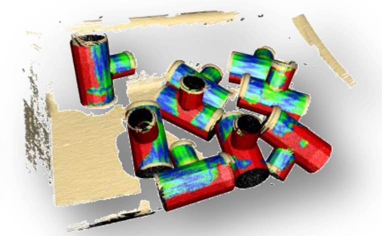
[Badami, Stückler, Behnke: SPME 2013]

Bin Picking

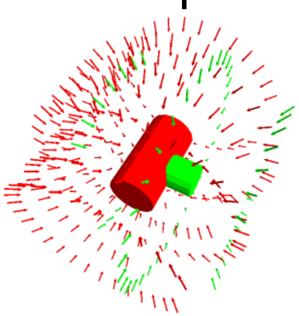
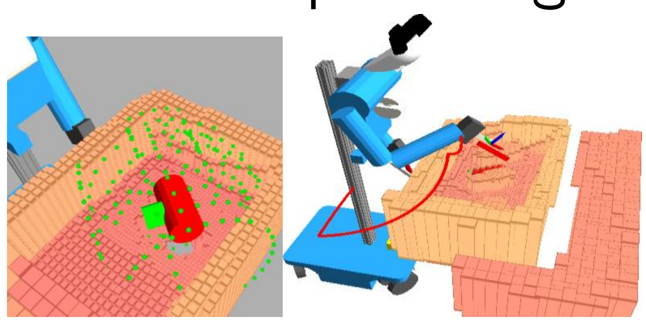
- Known objects in transport box


- Matching of graphs of 2D and 3D shape primitives



- Grasp and motion planning

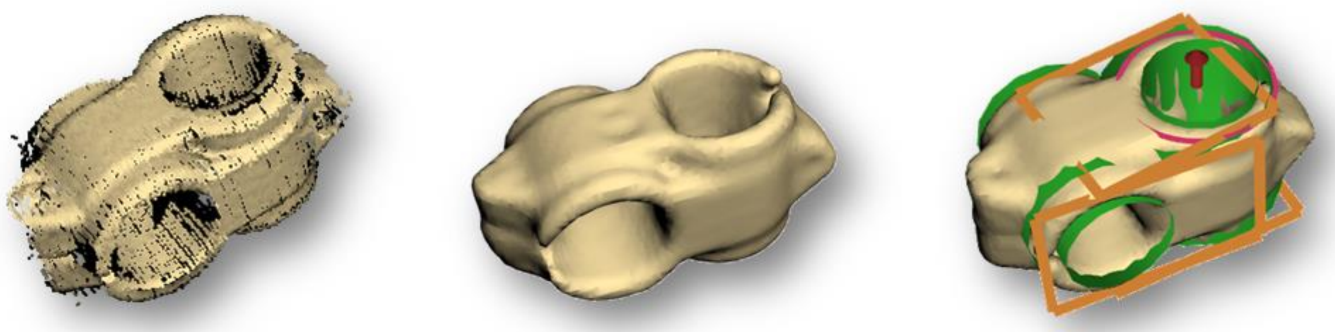
Offline

Online

[Nieuwenhuisen et al.: ICRA 2013]

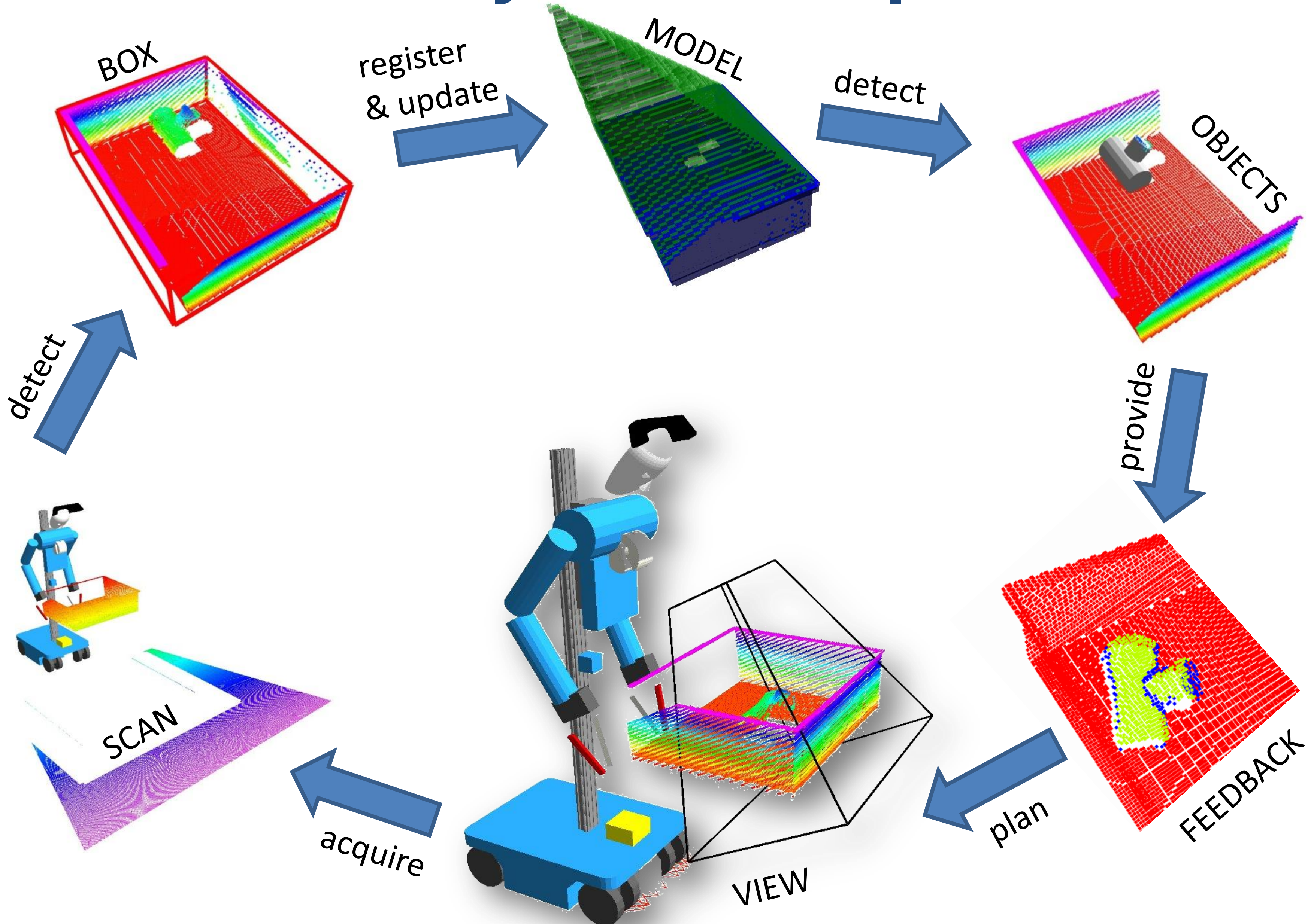
Learning of Object Models

- Scan multiple objects in different poses
- Find support plane and remove it
- Segment views
- Register views using ICP
- Recognize geometric primitives



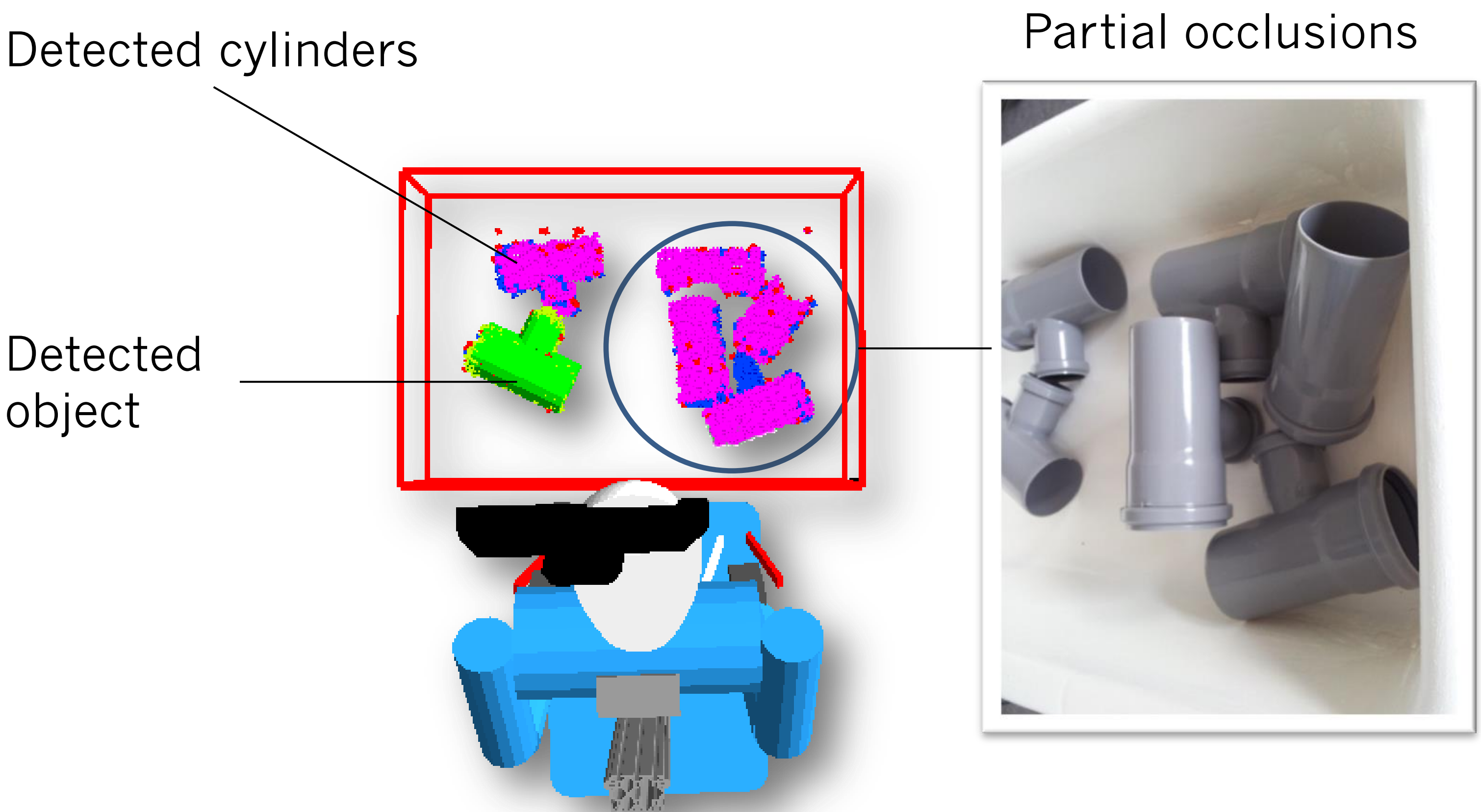
Registered views Surface reconstruction Detected primitives

Active Object Perception



[Holz et al. STAR 2014]

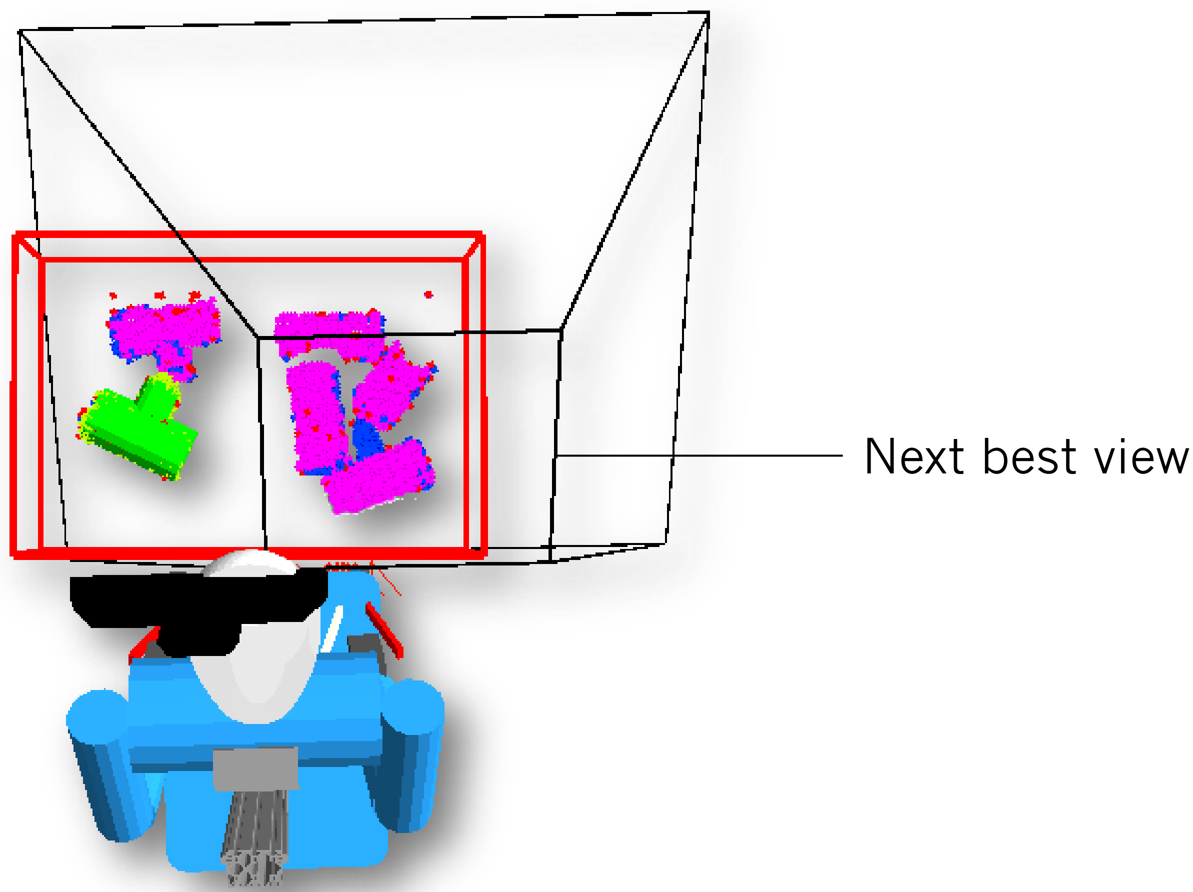
Active Object Perception



- Efficient exploration of the part arrangement in the transport boxes to handle occlusions

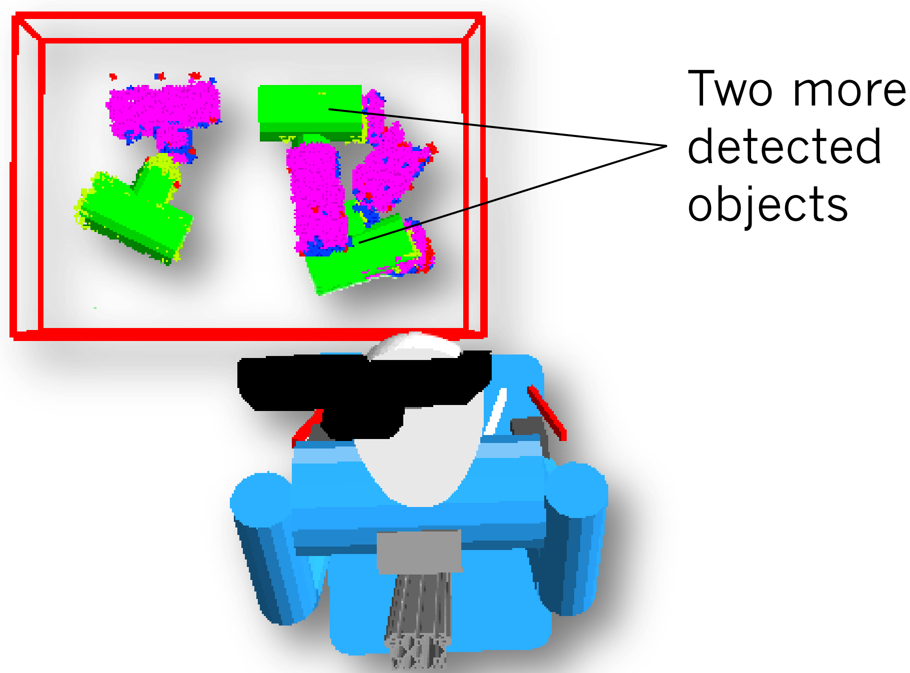
[Holz et al. STAR 2014]

Active Object Perception



- Efficient exploration of the part arrangement in the transport boxes to handle occlusions

Active Object Perception

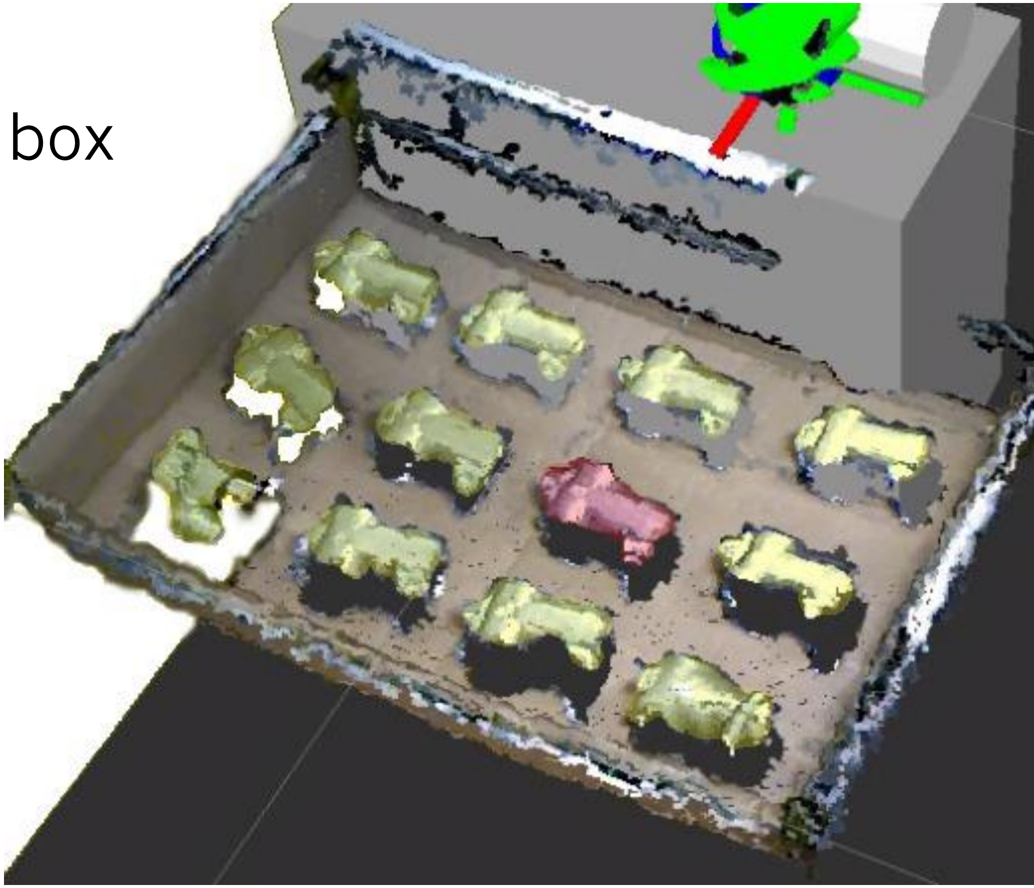


- Efficient exploration of the part arrangement in the transport boxes to handle occlusions

[Holz et al. STAR 2014]

Industrial Application: Depalettizing

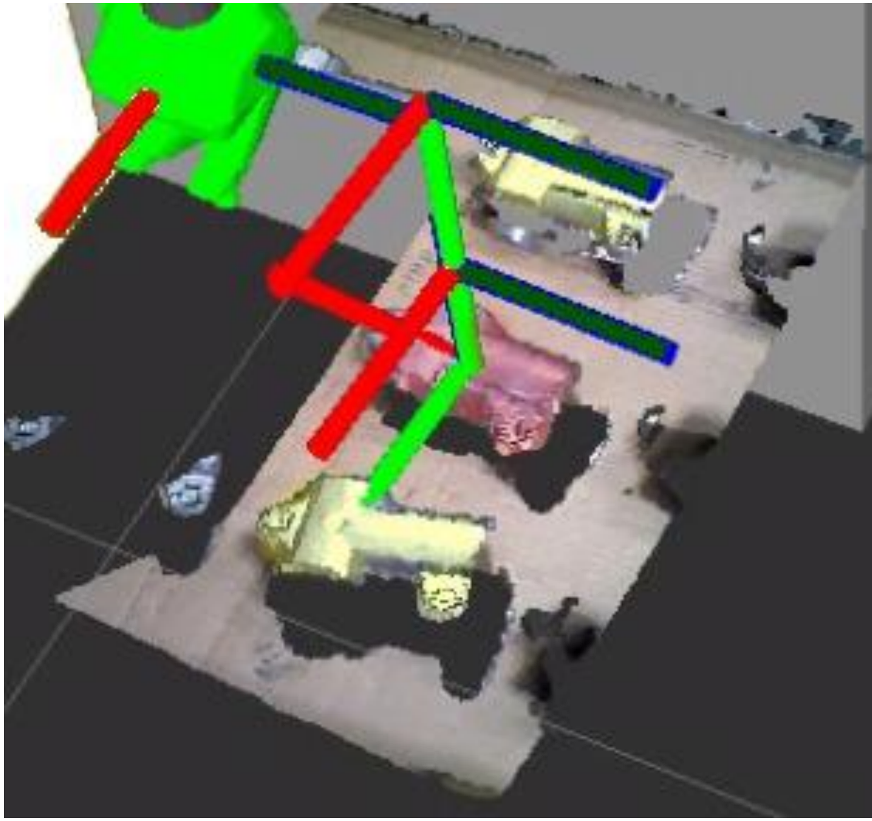
- Using work space RGB-D camera
- Initial pose of transport box roughly known
- Detect dominant horizontal plane above ground
- Cluster points above support plane
- Estimate main axes



[Holz et al. IROS 2015]

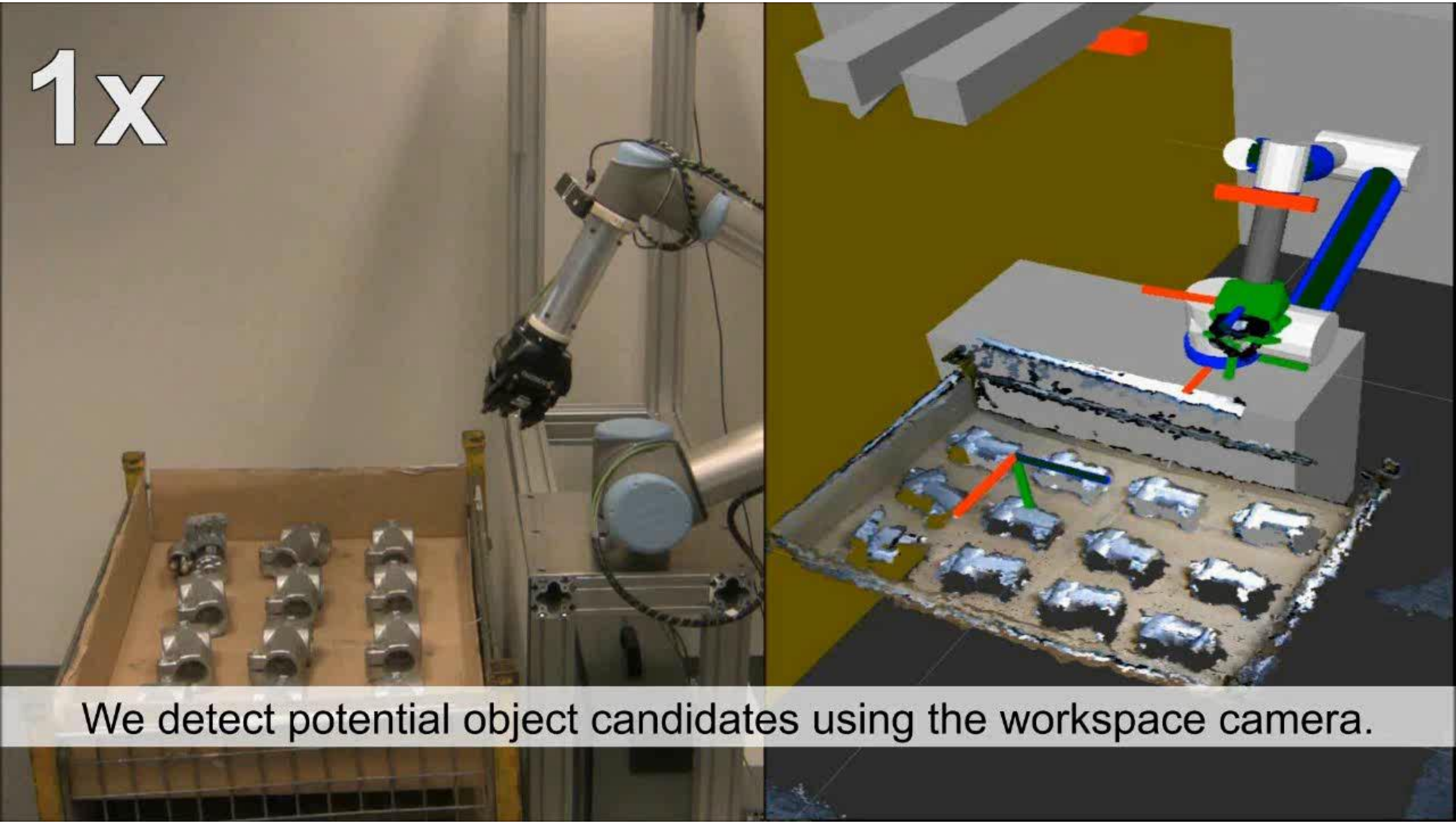
Object View Registration

- Wrist RGB-D camera moved above innermost object candidate
- Object views are represented as Multiresolution Surfel Map
- Registration of object view with current measurements using soft assignments
- Verification based on registration quality



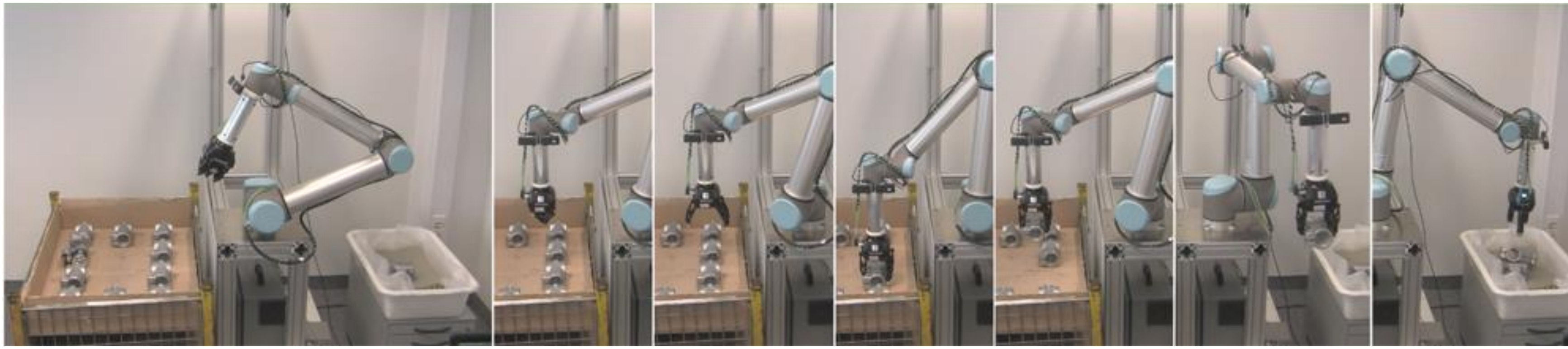
[Holz et al. IROS 2015]

Part Detection and Grasping



[Holz et al. IROS 2015]

Depalletizing Results: 10 Runs



- Total time

Component	Mean	Std	Min	Max
Object detection and grasping	13.84 s	1.89 s	10.42 s	23.81 s
Full cycle (incl. release and returning to initial pose)	34.57 s	3.01 s	29.53 s	49.52 s

- Component times and success rates

Component	Mean	Std	Min	Max	Success Rate
Initial object detection	26.3 ms	10.3 ms	0.02 ms	38.5 ms	100 %
Detecting that the pallet is empty					100 %
Object localization & verification	532.7 ms	98.2 ms	297.0 ms	800.1 ms	100 %
Identifying wrong objects					100 %
Grasping a found object	7.80 s	0.56 s	6.90 s	10.12 s	99 %

[Holz et al. IROS 2015]

Part Verification Results

- Parts used for verification

- Detection confidences

Object	Mean	Std	Min	Max
Correct object ("cross clamp")	0.901	0.024	0.853	0.951
Similar cross clamp (pose 1)	0	0	0	0
Similar cross clamp (pose 2)	0.407	0.034	0.299	0.452
Small starter	0	0	0	0
Large starter	0.505	0.055	0.398	0.581
Smaller cross clamp	0	0	0	0

Different Lighting Conditions

Artificial light and day light

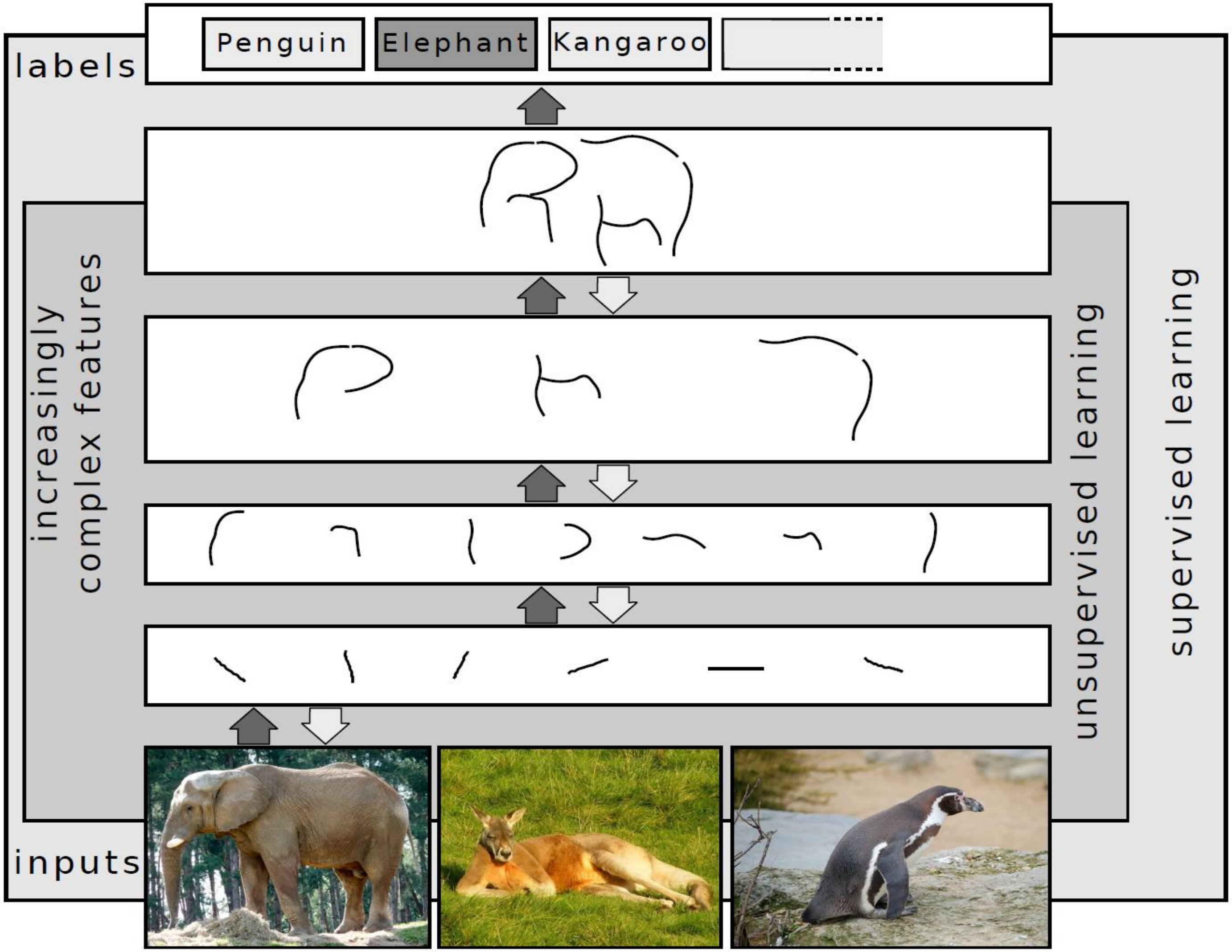
Only daylight

Low light

- In all cases, the palette was successfully cleared.

[Holz et al. IROS 2015]

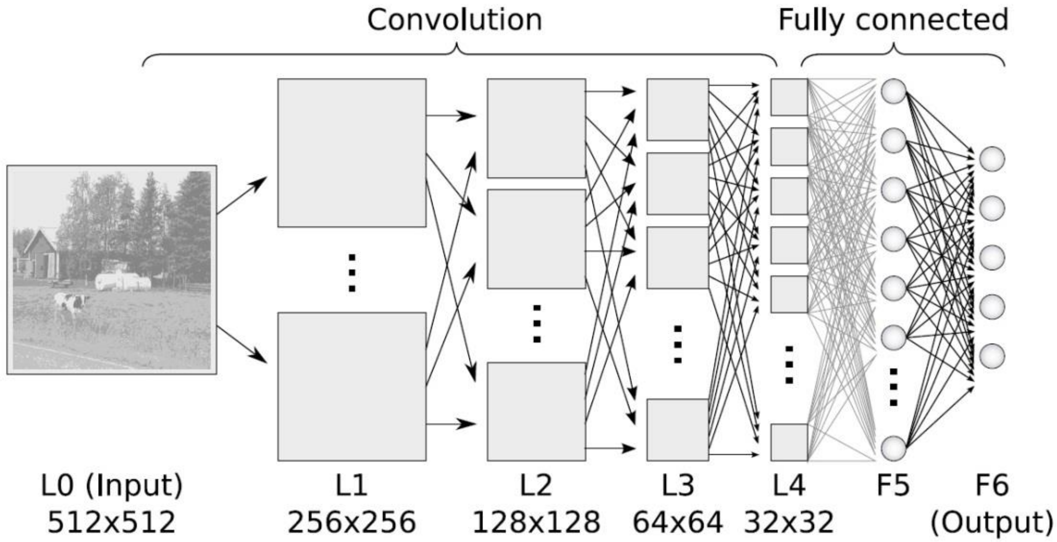
Deep Learning

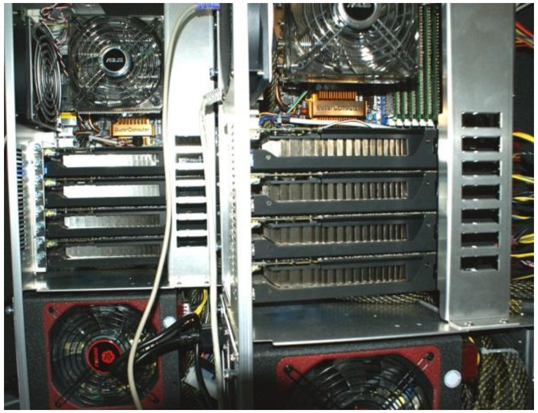


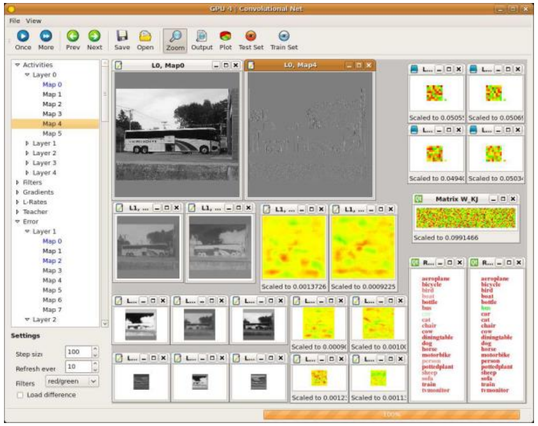
[Schulz and Behnke, KI 2012]

GPU Implementations (CUDA)

- Affordable parallel computers
- General-purpose programming
- Convolutional [Scherer & Behnke, 2009]



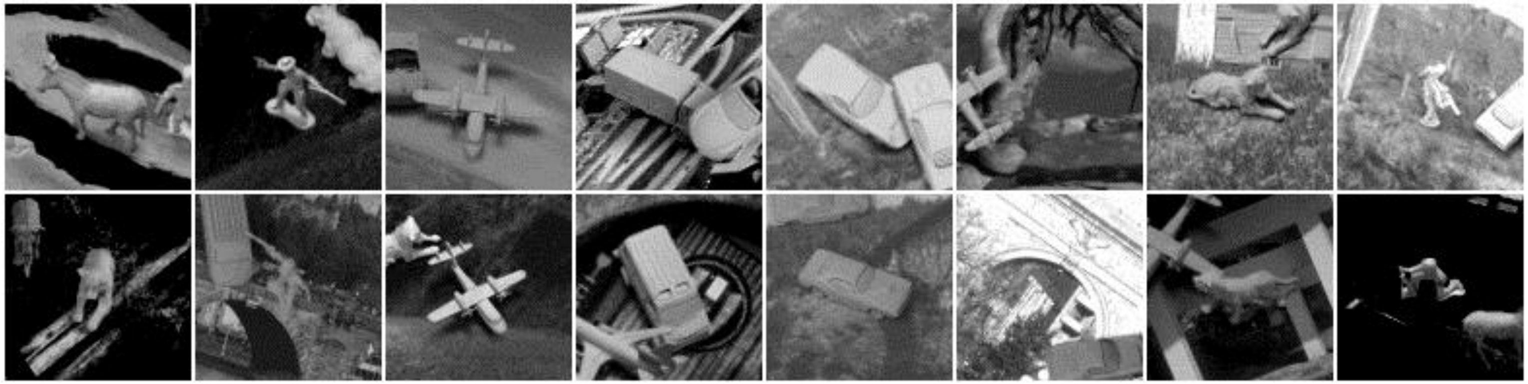




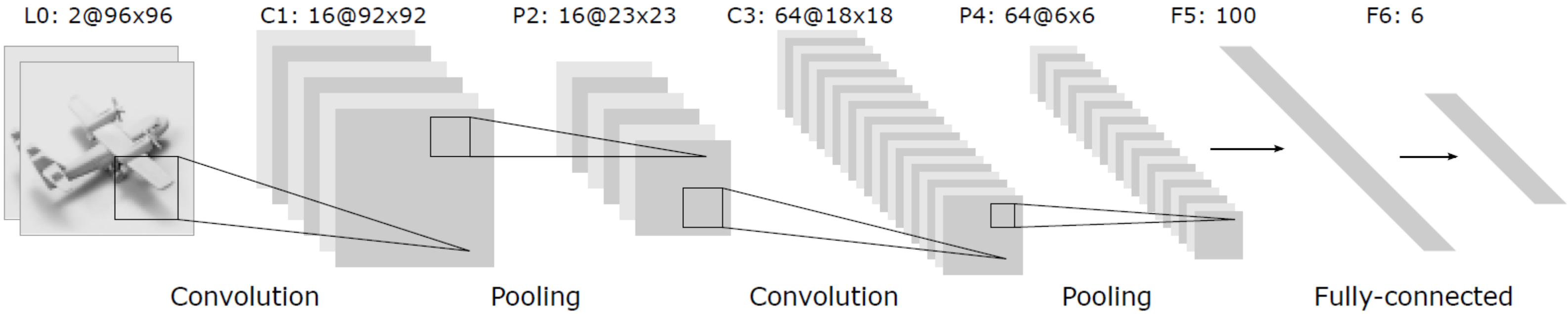
- Local connectivity [Uetz & Behnke, 2009]

Image Categorization: NORB

- 10 categories, jittered-cluttered



- **Max-Pooling**, cross-entropy training



- Test error: 5,6% (LeNet7: 7.8%)

[Scherer, Müller, Behnke, ICANN'10]

Image Categorization: LabelMe

- 50,000 color images (256x256)
- 12 classes + clutter (50%)

tree 1.0 | car 1.0 | building 1.0 | window 1.0 | person 1.0 | keyboard 1.0 | sign 1.0 | bookshelf 1.0

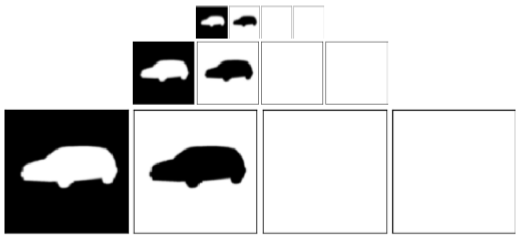
car 0.21 | person 0.54 | window 0.66 | building 1.0, tree 0.03 | (none) | (none) | (none) | (none)

- Error TRN: 3.77%; TST: 16.27%
- Recall: 1,356 images/s

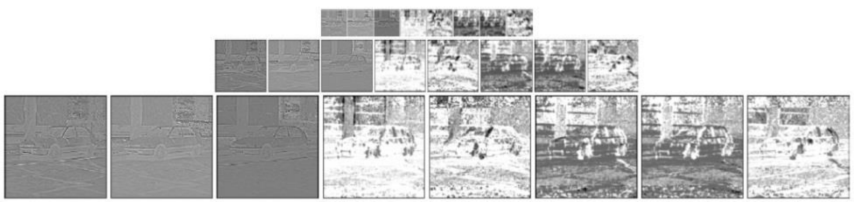
[Uetz, Behnke, ICIS2009]

Object-class Segmentation

- Class annotation per pixel

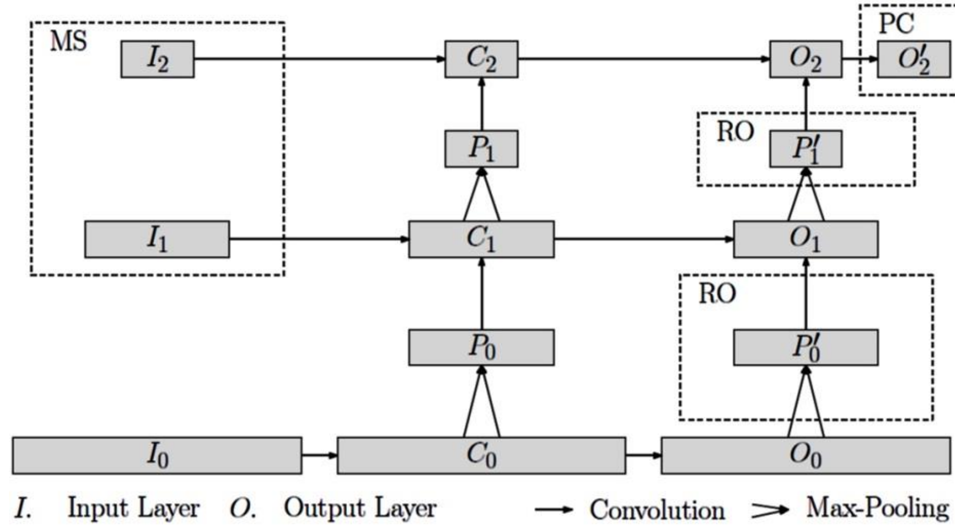


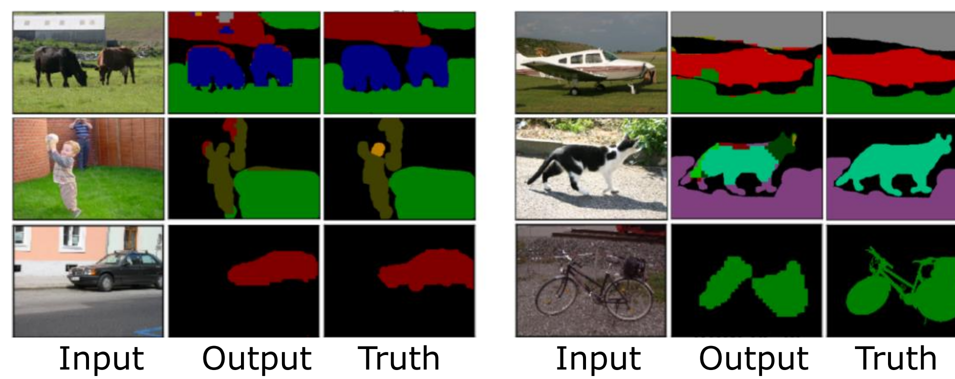
- Multi-scale input channels



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

[Schulz, Behnke 2012]





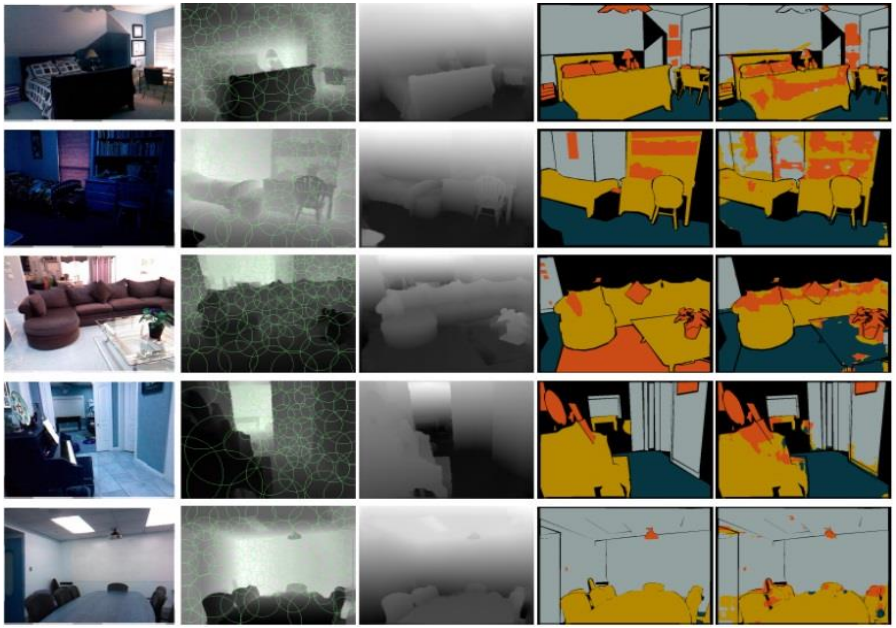
Object Detection in Images

- 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

- Scale input according to depth
- Compute pixel height



RGB Depth Height Truth Output

NYU Depth V2

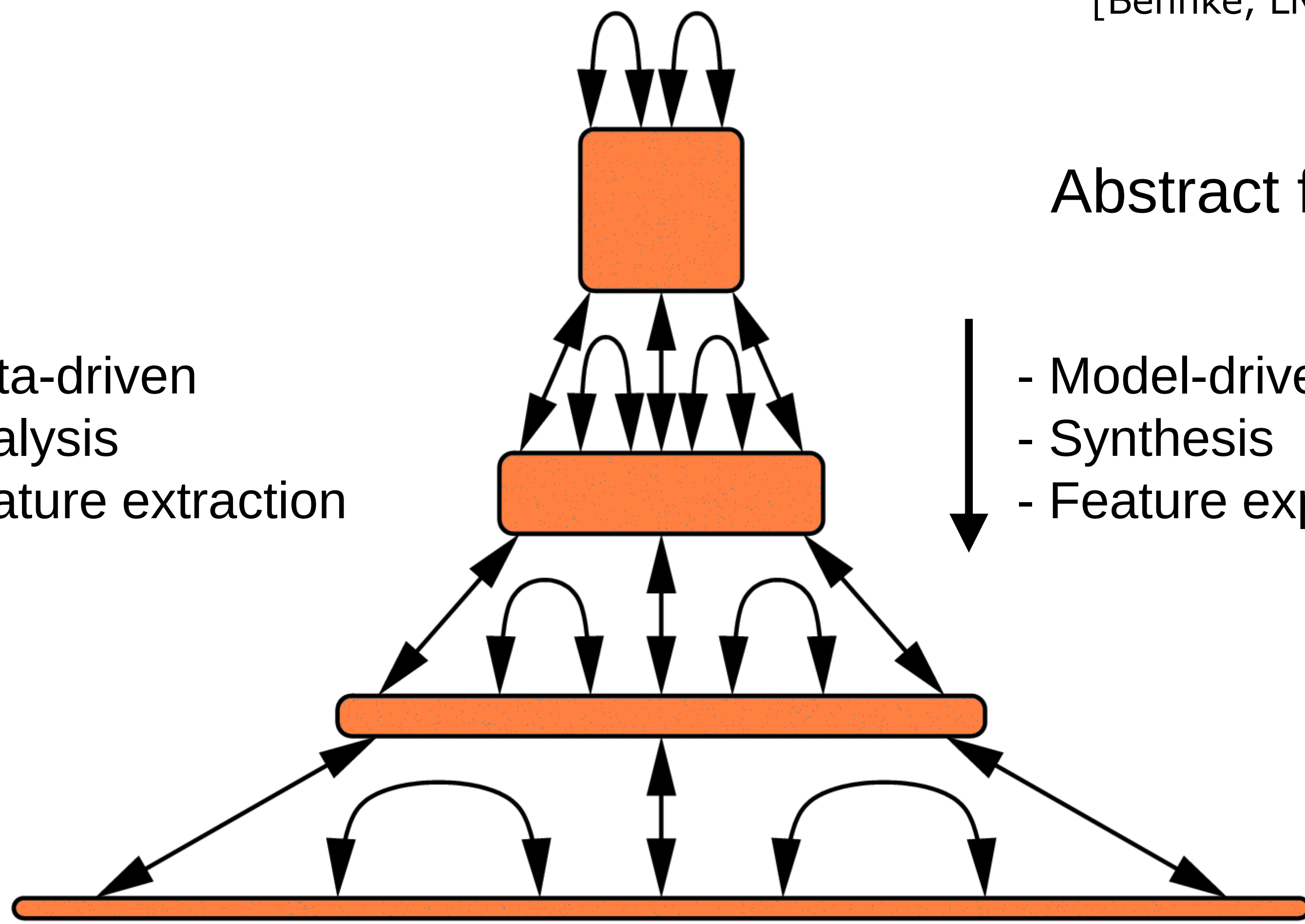
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]	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]

Neural Abstraction Pyramid

[Behnke, LNCS 2766, 2003]



Abstract features

- Data-driven
- Analysis
- Feature extraction

- Model-driven
- Synthesis
- Feature expansion

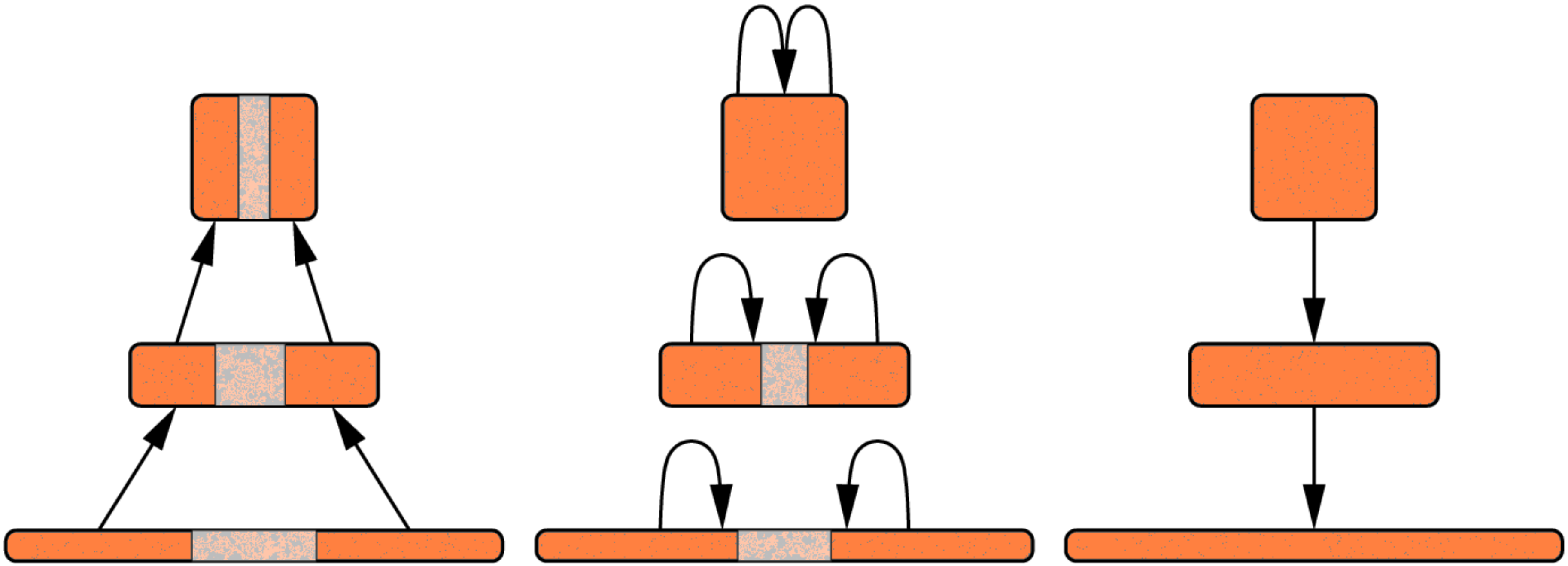
Signals

- Grouping - Competition - Completion

Iterative Interpretation

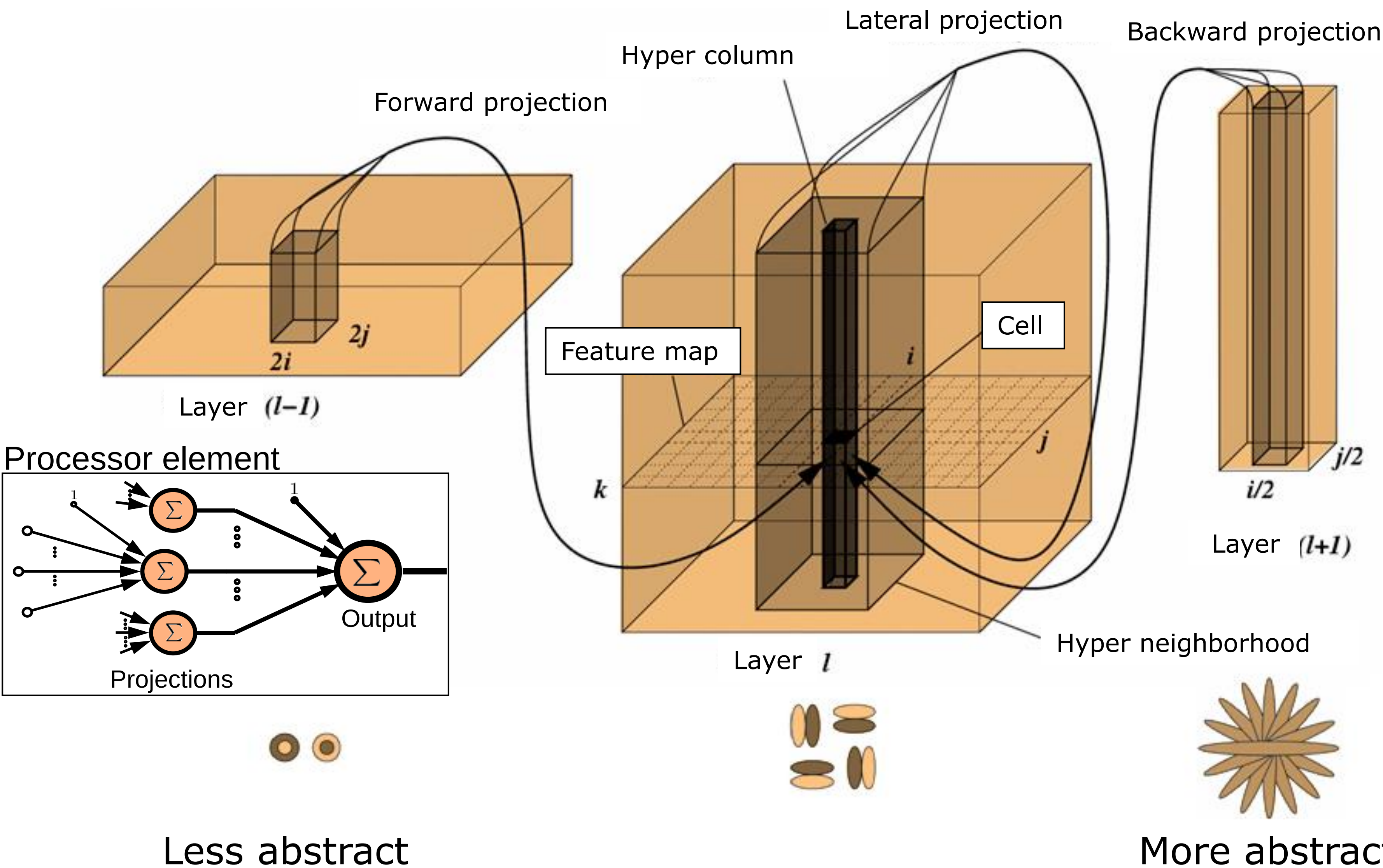
[Behnke, LNCS 2766, 2003]

- Interpret most obvious parts first



- Use partial interpretation as context to resolve local ambiguities

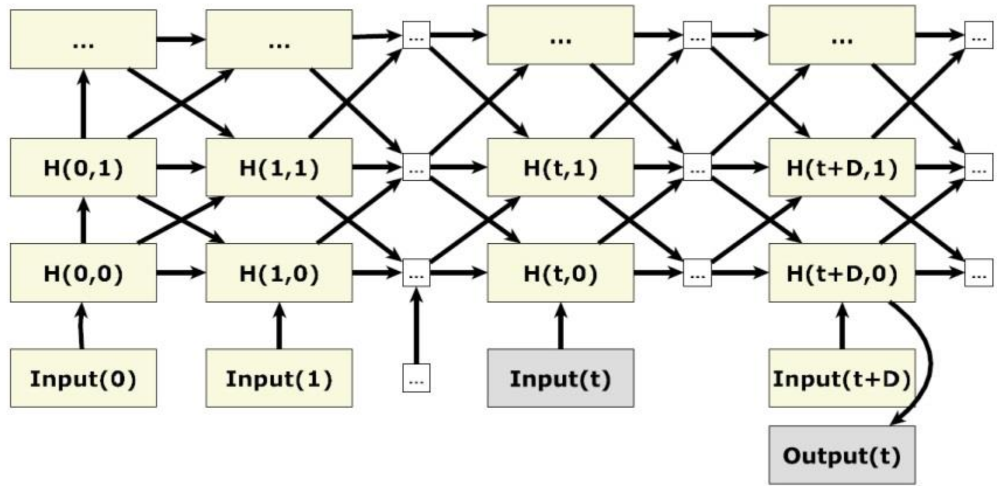
Local Recurrent Connectivity

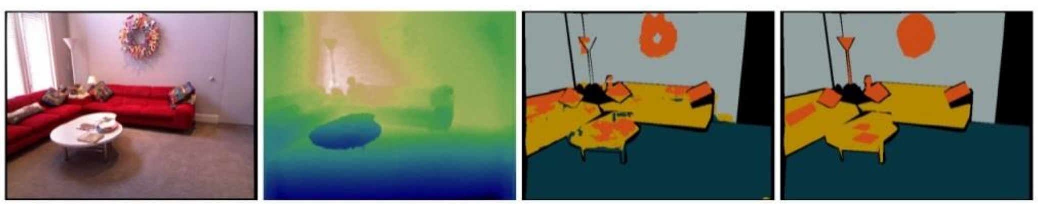


[Behnke, LNCS 2766, 2003]

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

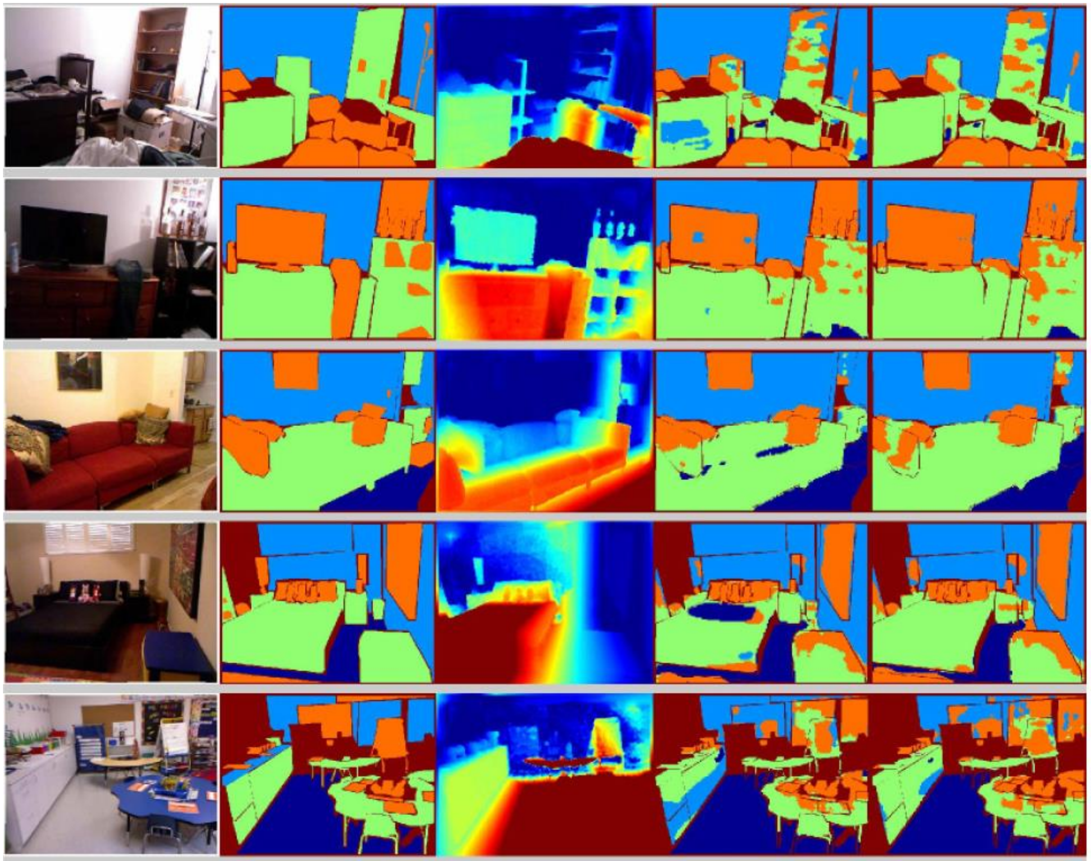
Method	Class Accuracies (%) ground	struct	furnit	prop	Average (%) Class	Pixel
Höft *et al.* [19]	77.9	65.4	55.9	49.9	62.0	61.1
Unidirectional + MS	73.4	66.8	**60.3**	49.2	62.4	63.1
Schulz *et al.* [20] (no height)	87.7	70.8	57.0	53.6	67.3	65.5
Unidirectional + SW	**90.0**	**76.3**	52.1	**61.2**	**69.9**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

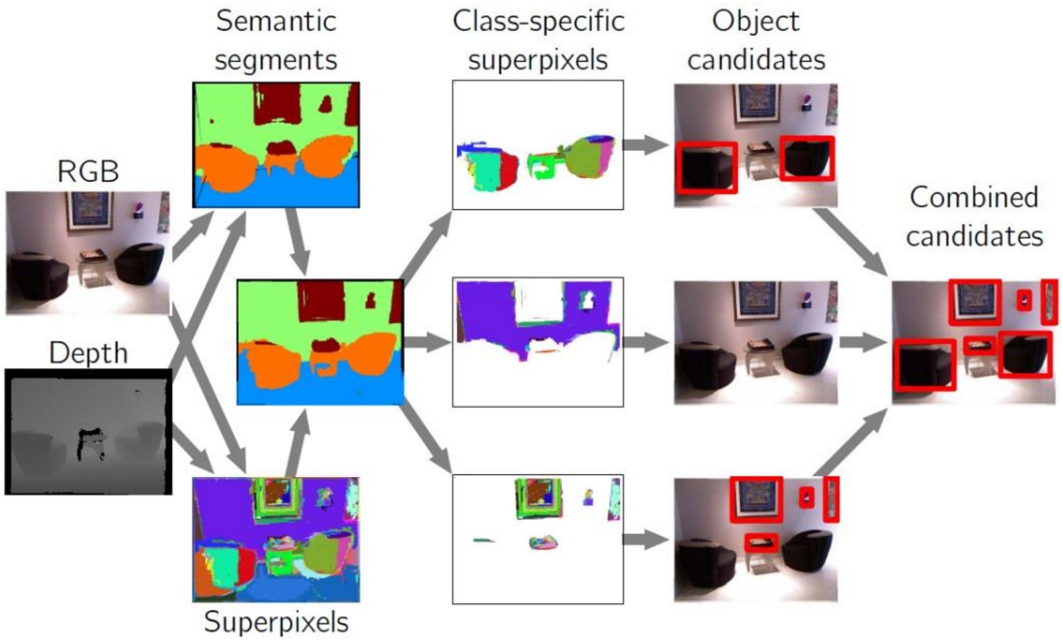
[Husain et al. under review]



RGB Truth DistWall OutWO OutWithDist

Semantic Segmentation Priors for Object Discovery

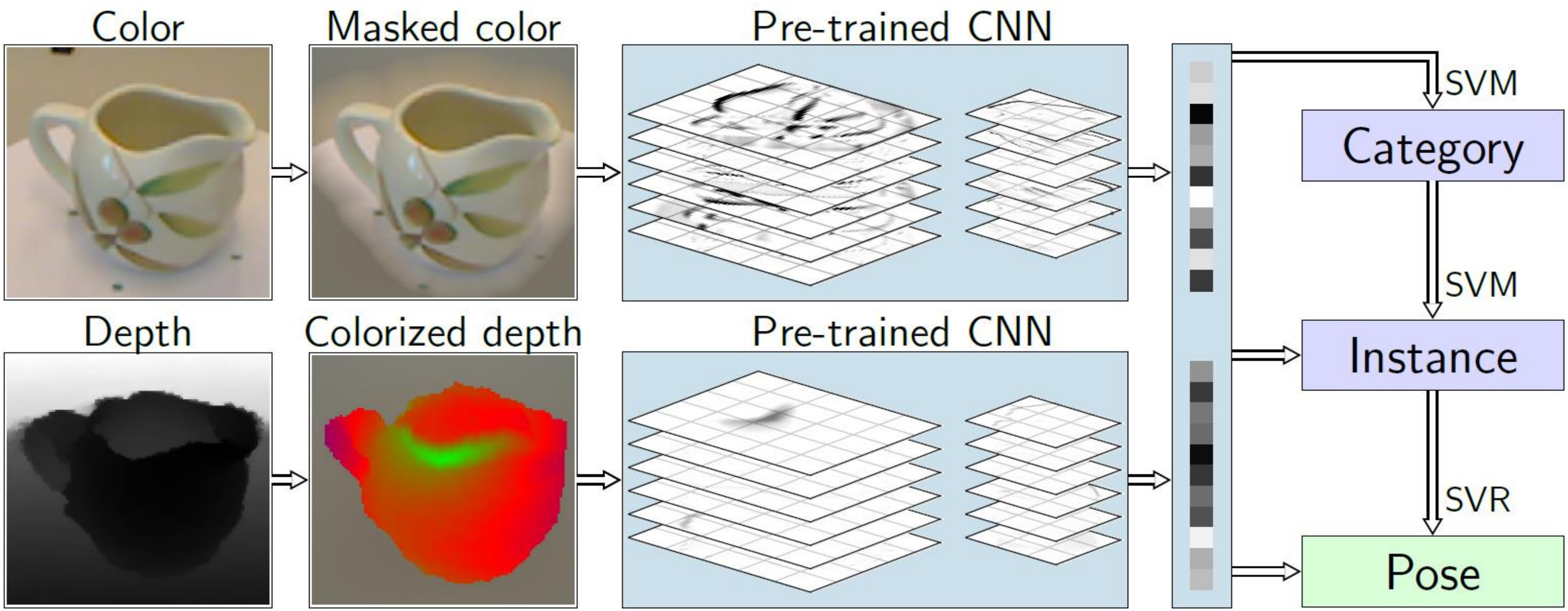
- Combine bottom-up object discovery and semantic priors
- Semantic segmentation used to classify color and depth superpixels
- Higher recall, more precise object borders



[Garcia et al. under review]

RGB-D Object Recognition and Pose Estimation

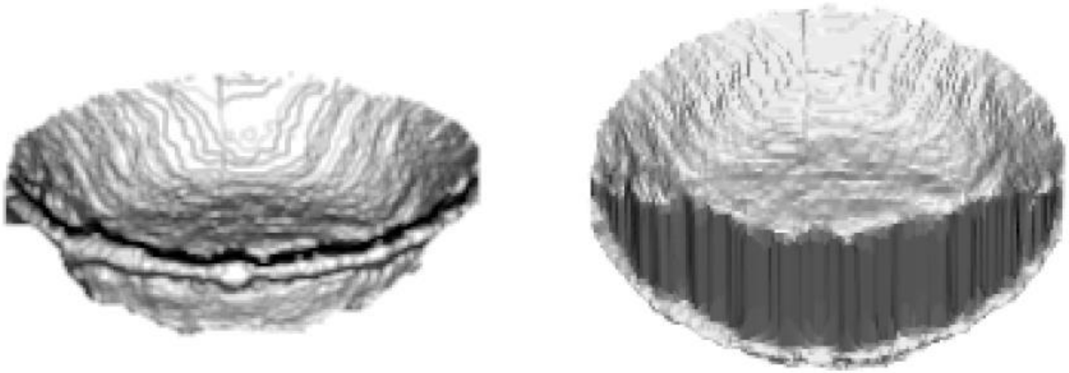
- Use pretrained features from ImageNet



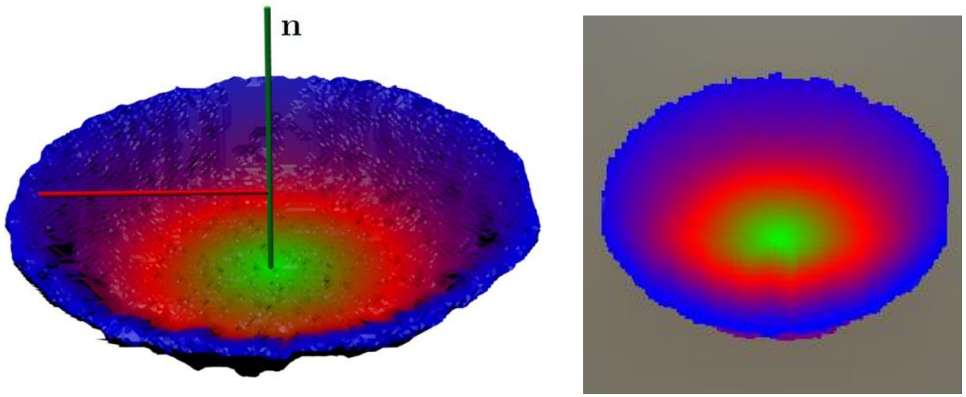
[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view



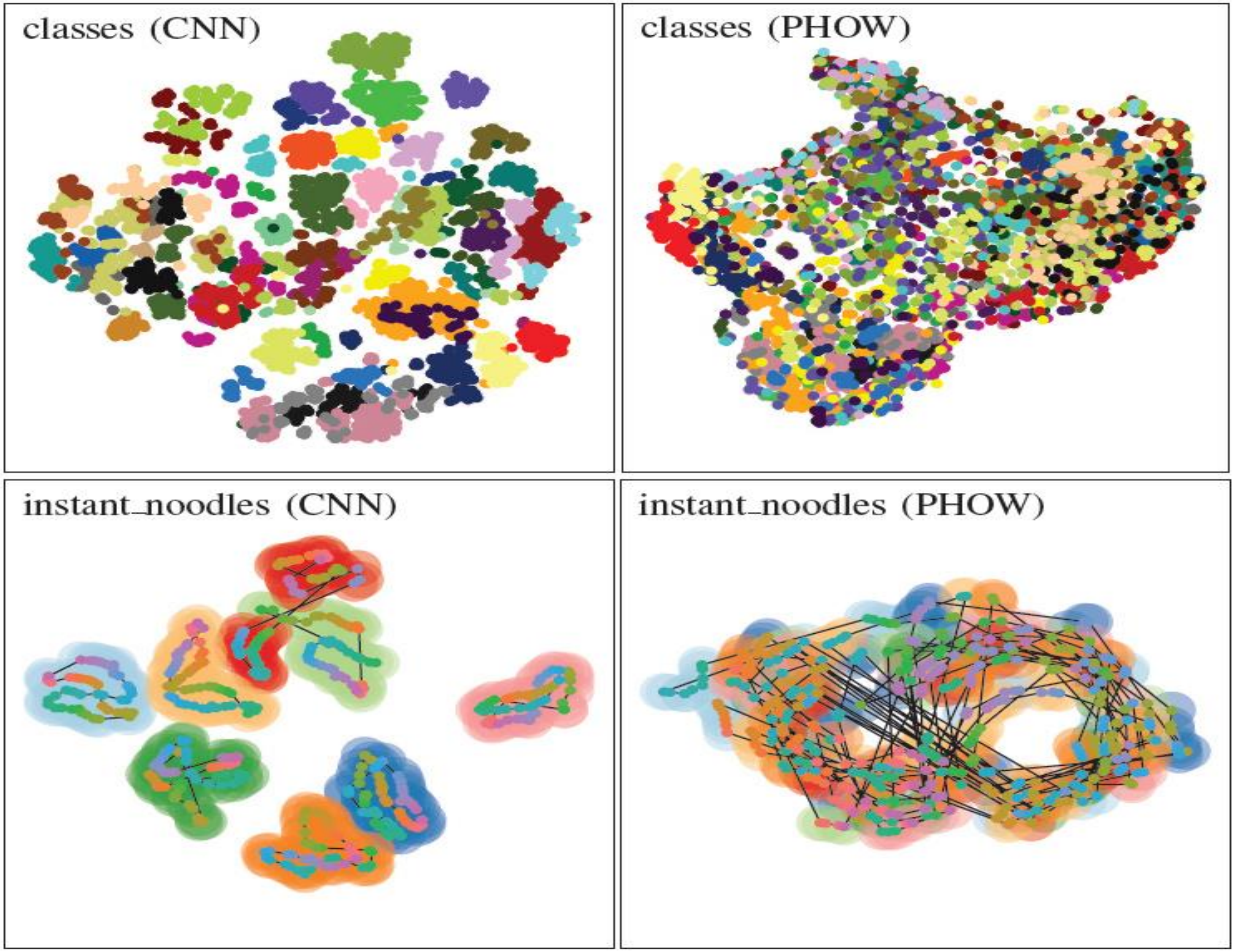
- Colorization based on distance from center vertical



[Schwarz, Schulz, Behnke, ICRA2015]

Features Disentangle Data

- t-SNE embedding



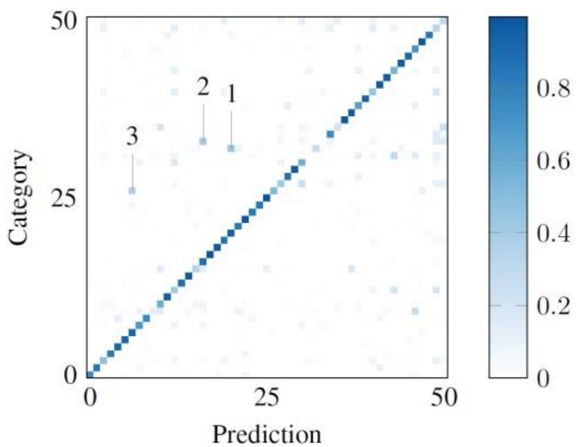
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

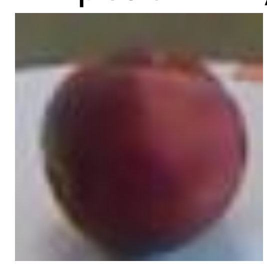
Method	Category Accuracy (%) RGB	Category Accuracy (%) RGB-D	Instance Accuracy (%) RGB	Instance Accuracy (%) RGB-D
Lai *et al.* [1]	74.3 ± 3.3	81.9 ± 2.8	59.3	73.9
Bo *et al.* [2]	82.4 ± 3.1	87.5 ± 2.9	**92.1**	92.8
PHOW[3]	80.2 ± 1.8	—	62.8	—
Ours	$\mathbf{83.1 \pm 2.0}$	88.3 ± 1.5	92.0	**94.1**
Ours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

- Confusion



1: pitcher / coffe mug


2: peach / sponge




[Schwarz, Schulz, Behnke, ICRA2015]

Conclusion

- Semantic perception in everyday environments is challenging
- Simple methods rely on strong assumptions (e.g. support plane)
- 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, ...

Thanks for your attention!

Questions?