Learning Semantic Perception for Cognitive Robots

Sven Behnke

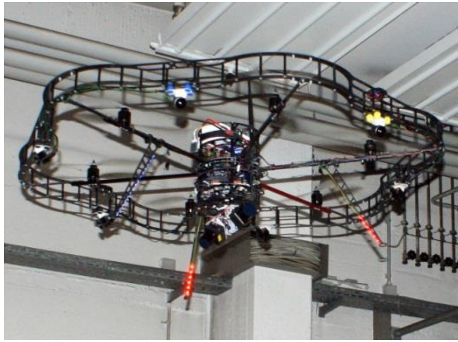
University of Bonn, Germany

Computer Science Institute VI

Autonomous Intelligent Systems

Some of Our Cognitive Robots

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



MAV

Soccer robot

Service robot

Exploration robot

Picking robot

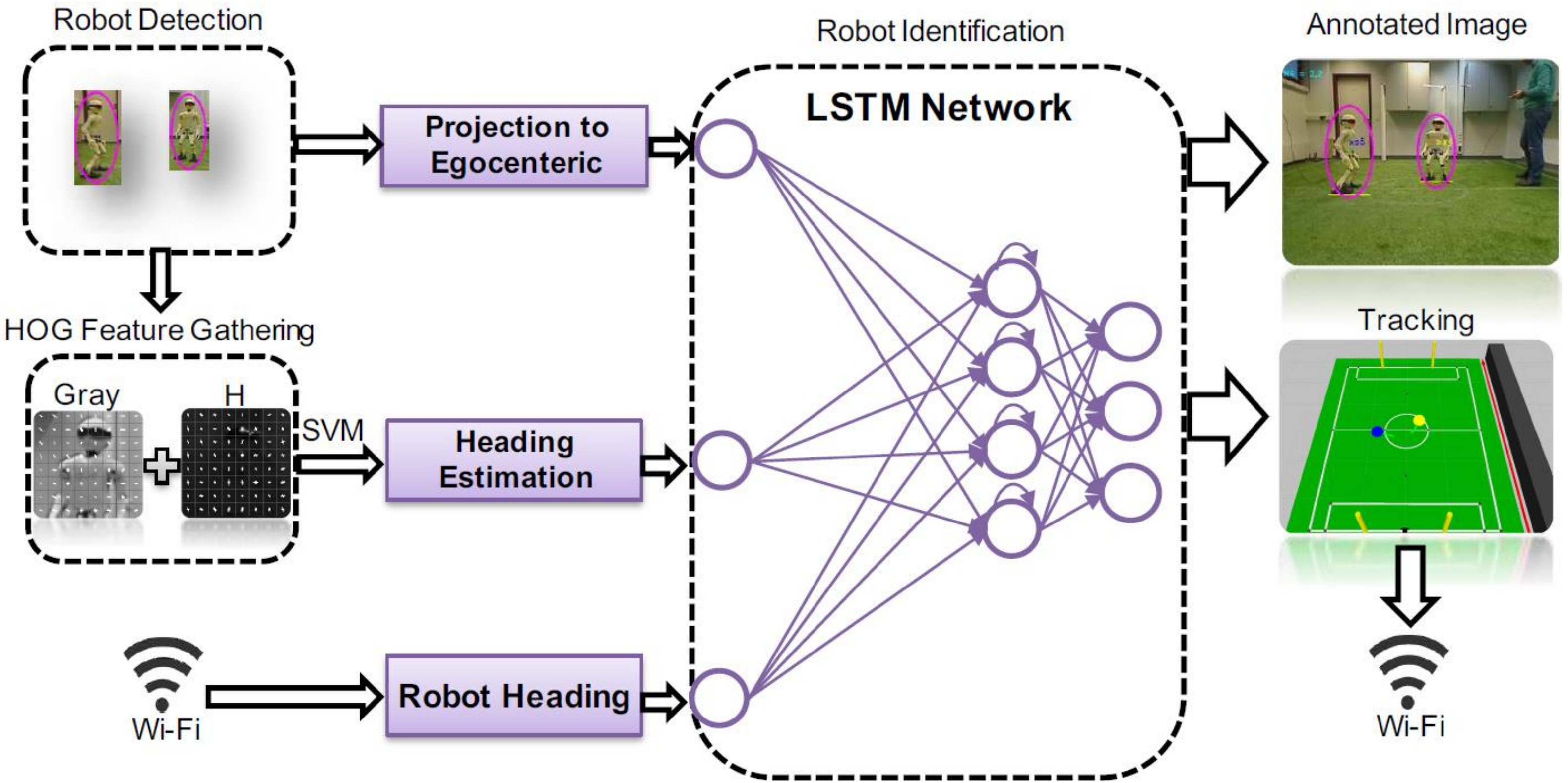
Visual Perception of Soccer Scene

[Farazi & Behnke, RoboCup 2016]

RoboCup 2016 TeenSize Final

universitätbonn ais

Robot Detection, Tracking & Identification

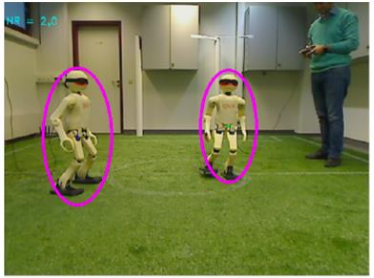


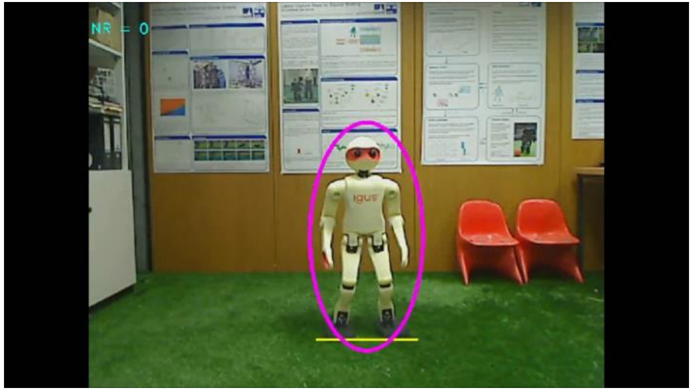
[Farazi & Behnke, IROS 2017]

Robot Detection

- Based on HoG features

- Scan line feet estimation

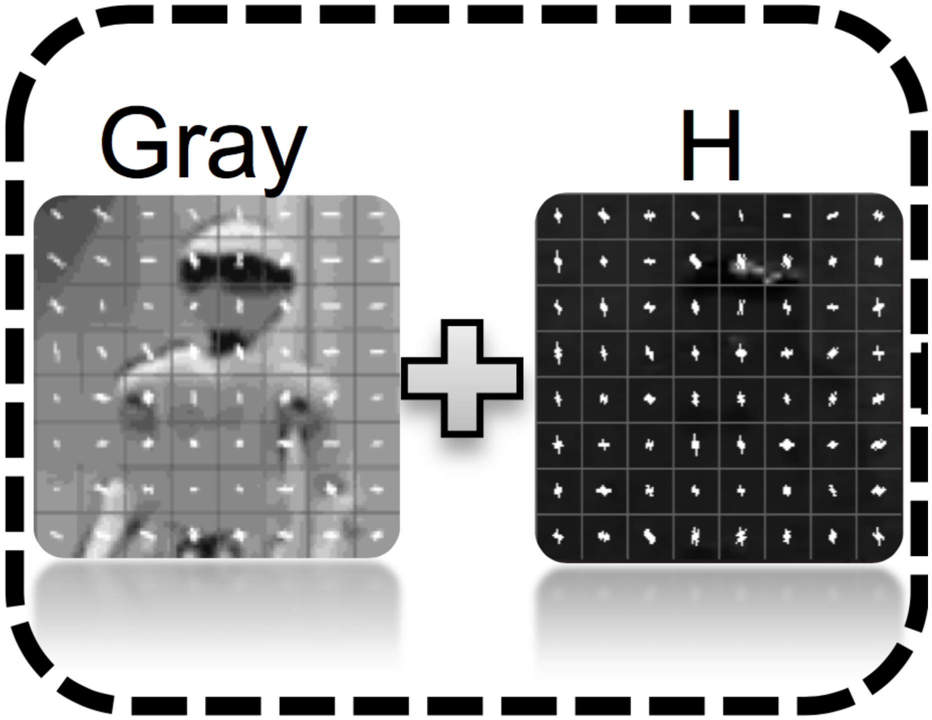




[Farazi & Behnke, IROS 2017]

Visual Heading Estimation

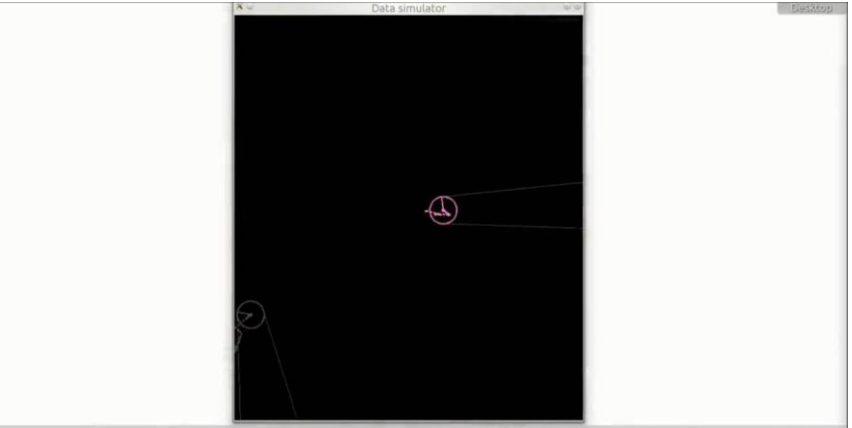
- Dense HOG on upper half of detection
- SVM multiclass classifier
- 10 classes (36° each)



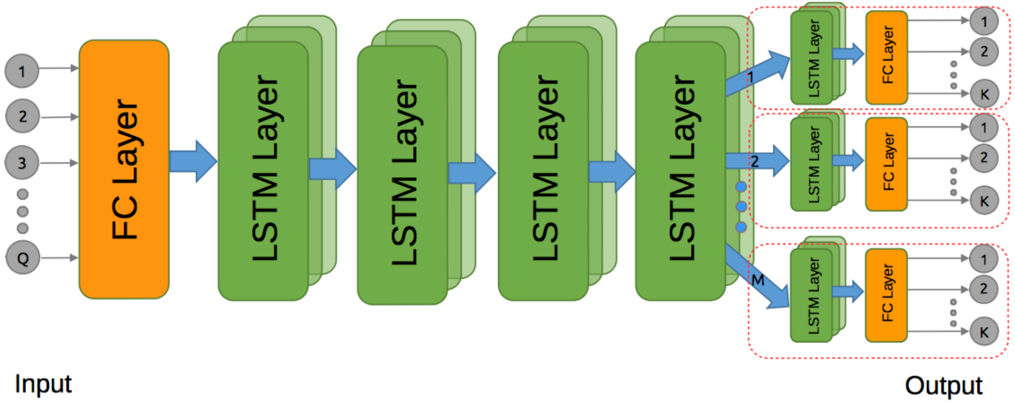
[Farazi & Behnke, IROS 2017]

Learning Data Association

- Recurrent neural network
- Training with simulated data



2D simulator for two robots and three detections



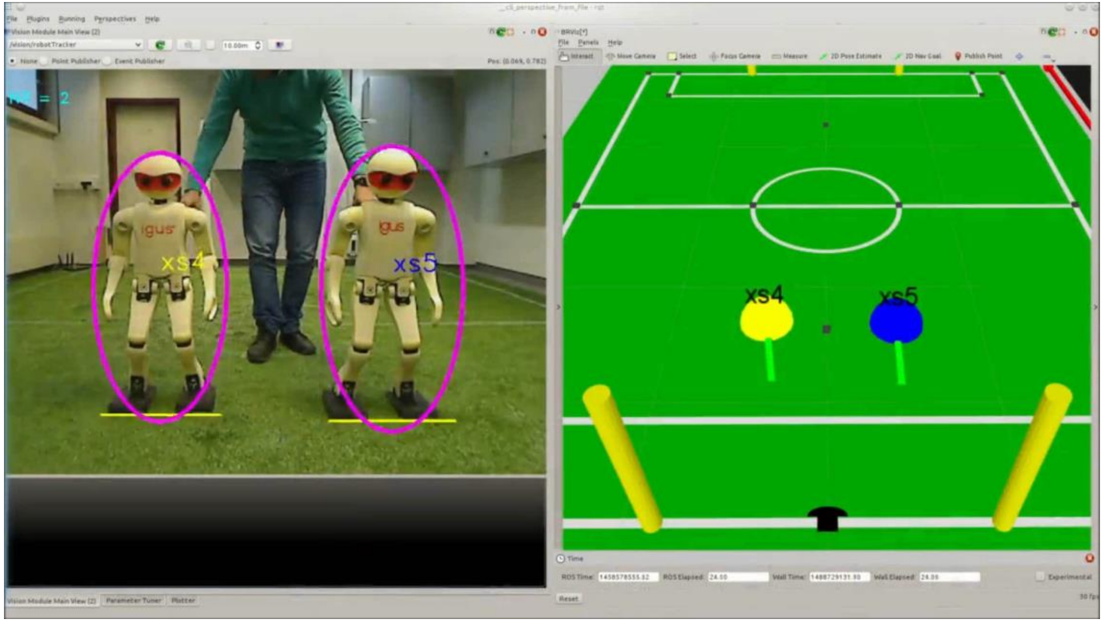
Setup	M=3, N=2	M=5, N=3	M=10, N=7
Human	94.3%	86.3%	**67.3%**
Kalman-HA	75.6%	72.2%	53.1%
ours	**96.2%**	**87.1%**	66.5%

- Fine-tuning on real data

[Farazi & Behnke, IROS 2017]

Real-Robot Experiment

- Three Igus humanoid robots, observer in goal area
- Randomly chosen sequences, 3140 frames in total
- Partial, short term and long term occlusions, Single forward 4ms (≈250Hz)



Baseline	Kalman-HA	Kalman-HA2	JPDA	**Ours**
Average error	0.67 m	0.30 m	0.29 m	**0.22 m**

Frames	200	400	800	Total
Kalman-HA	73.2%	75.5%	72.1%	73.8%
Kalman-HA2	87.2%	84.0%	86.3%	85.5%
JPDA	87.1%	84.6%	85.6%	86.3%
Deep LSTM (ours)	**89.8%**	**90.3%**	**92.4%**	**91.1%**

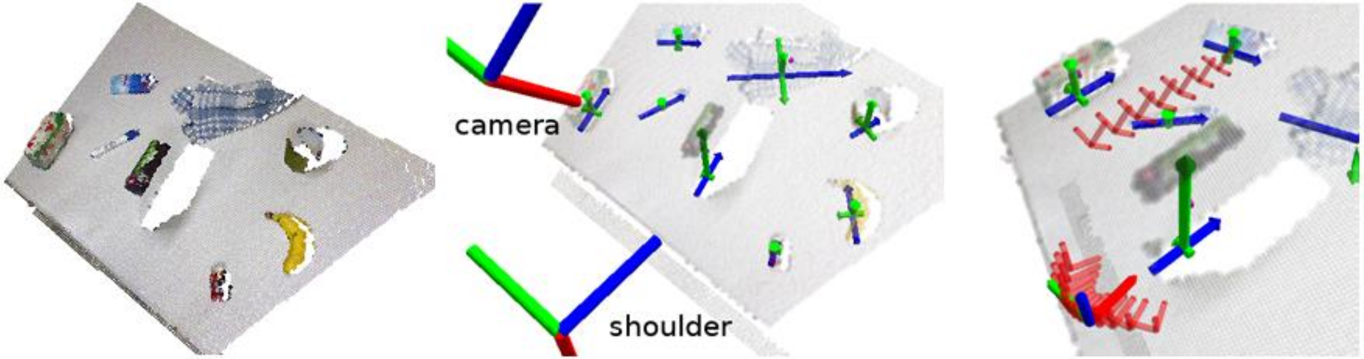
RoboCup 2017 AdultSize Final

Cognitive Service Robot Cosero

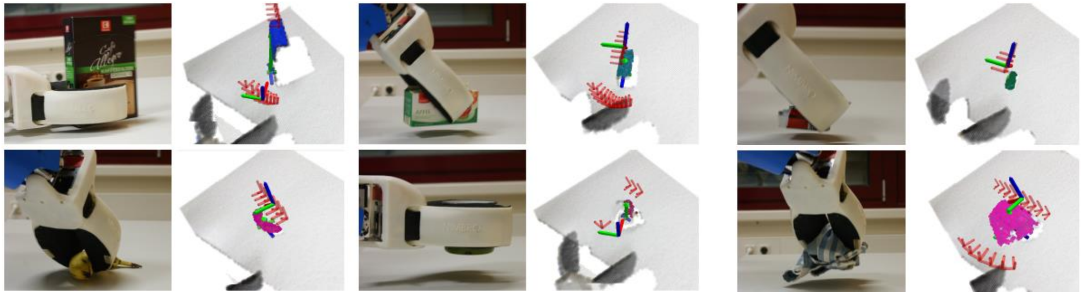
universitätbonn ais

Table-top Analysis and Grasp Planning

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



- Flexible grasping of many unknown objects

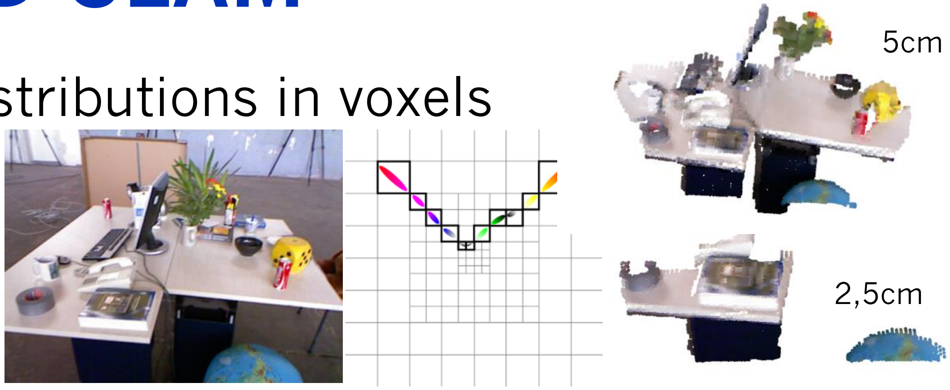


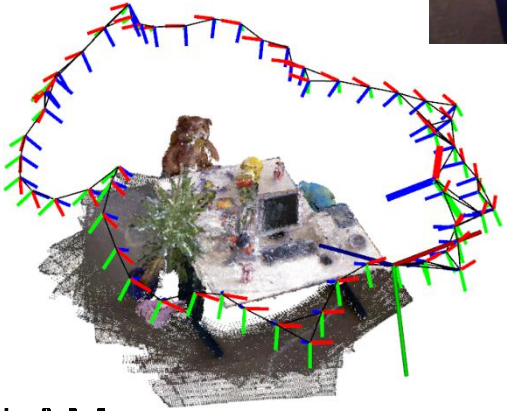
[Stückler et al, Robotics and Autonomous Systems, 2013]

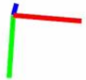
3D Mapping by RGB-D SLAM

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

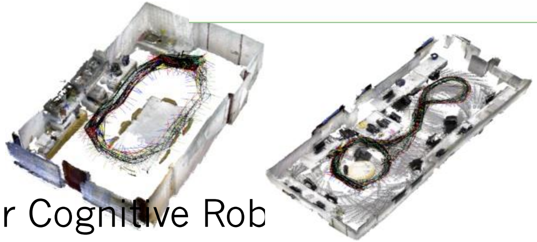
- Modelling of shape and color distributions in voxels
- Local multiresolution
- Efficient registration of views on CPU
- Global optimization







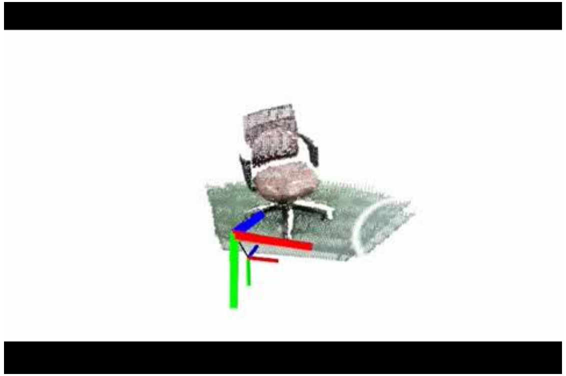
- Multi-camera SLAM



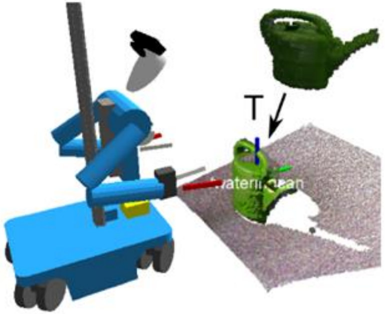
[Stoucken]

Learning and Tracking Object Models

- Modeling of objects by RGB-D-SLAM

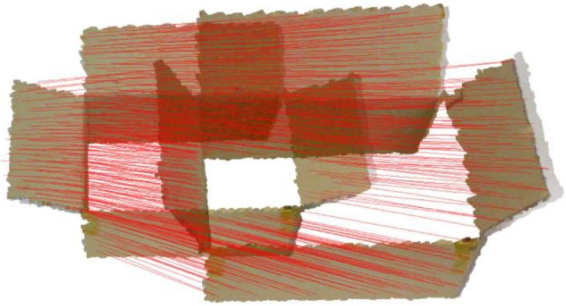


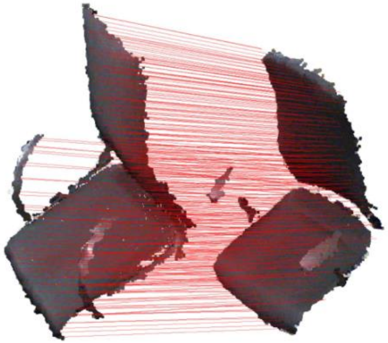
- Real-time registration with current RGB-D frame

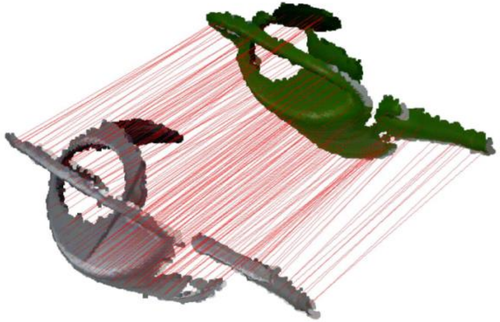


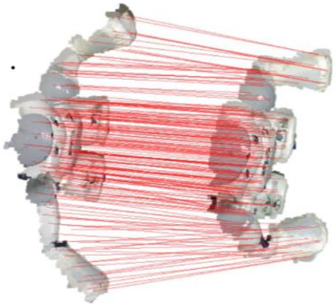
Deformable RGB-D-Registration

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



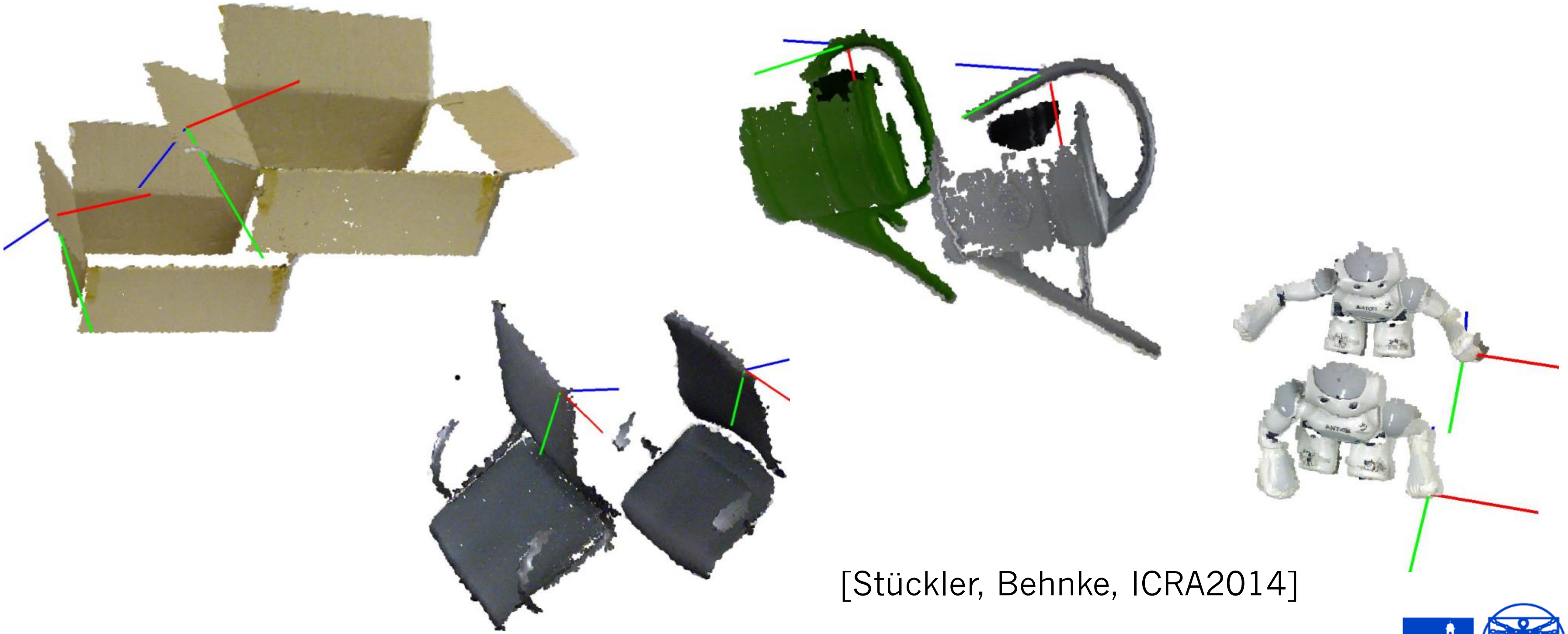






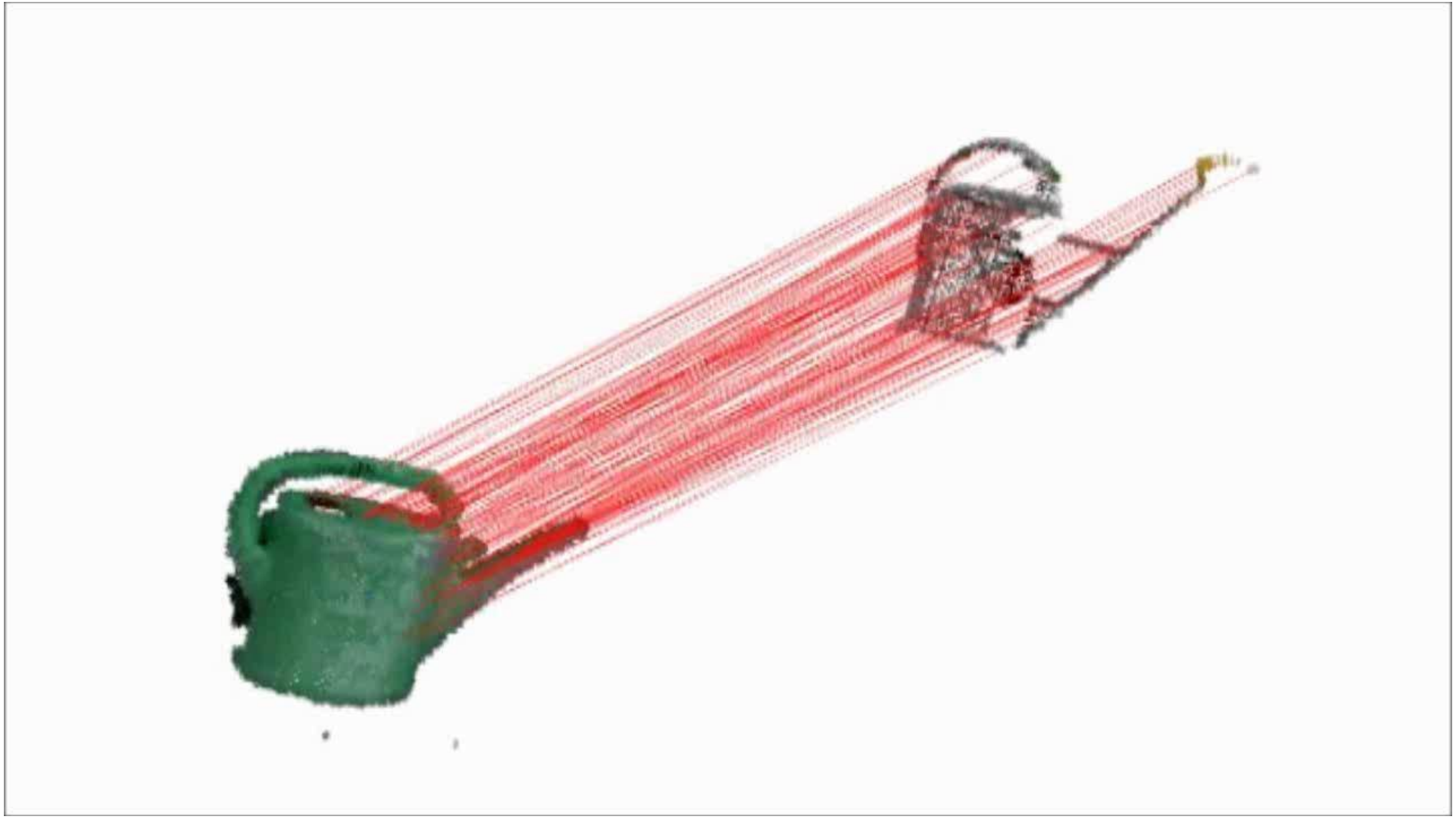
Transformation of Poses on Object

- Derived from the deformation field



[Stückler, Behnke, ICRA2014]

Grasp & Motion Skill Transfer



[Stückler, Behnke, ICRA2014]

Tool use: Bottle Opener

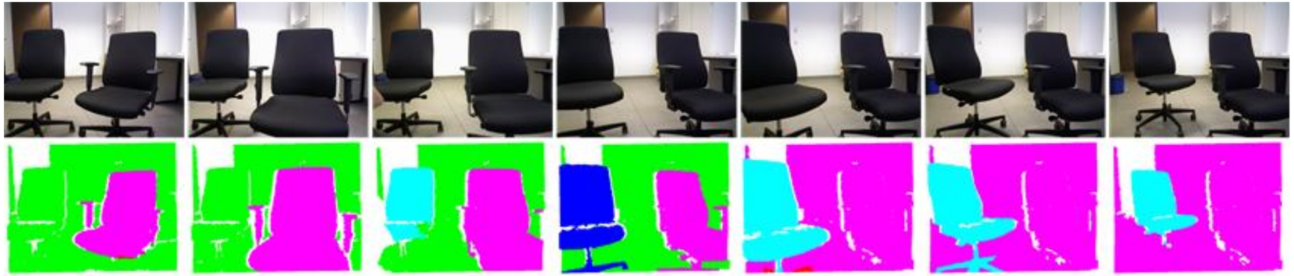
- Tool tip perception

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

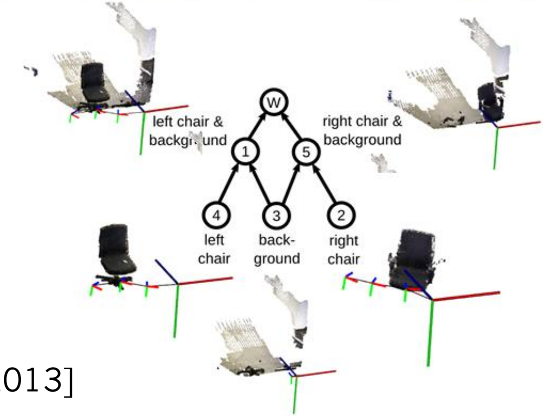
[Stückler, Behnke, Humanoids 2014]

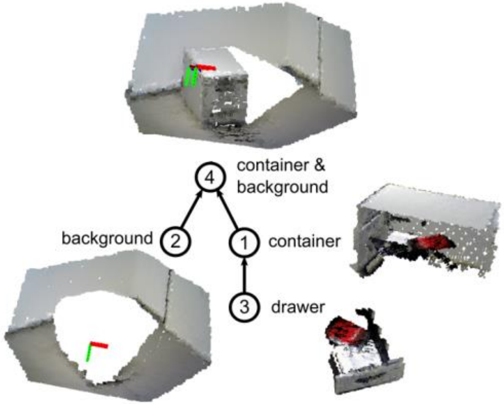
Hierarchical Object Discovery trough Motion Segmentation

- Simultaneous object modeling and motion segmentation



- Inference of a segment hierarchy

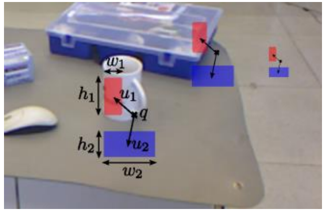


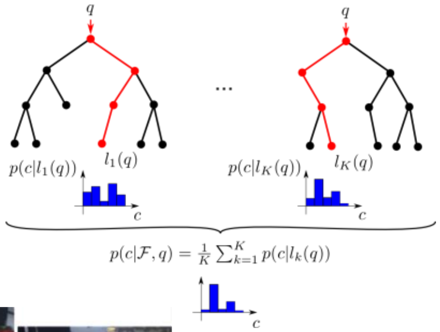


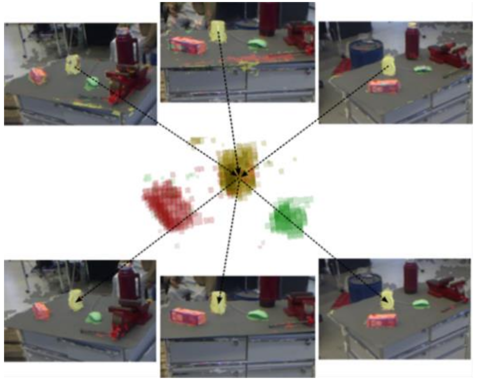
[Stückler, Behnke: IJCAI 2013]

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

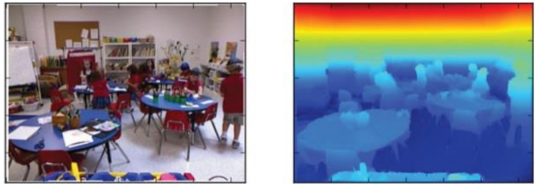
Segmentation

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**	

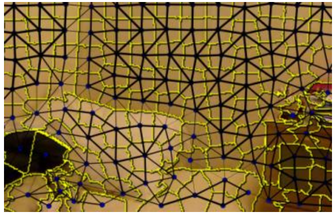
universitätbonn ais

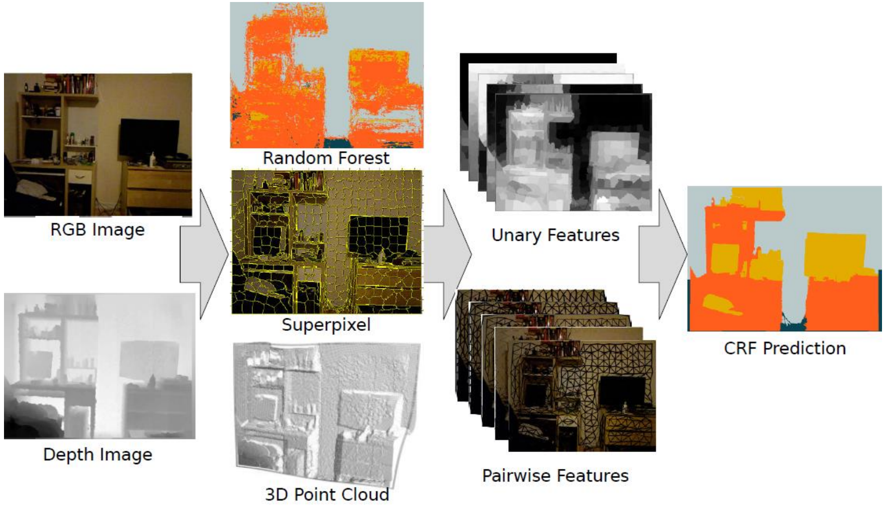
Learning Depth-sensitive CRFs

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



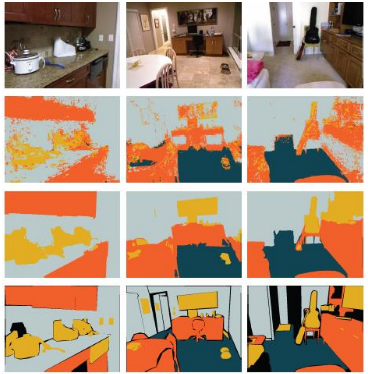
- Pairwise features
  - Color contrast
  - Vertical alignment
  - Depth difference
  - Normal differences





- 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

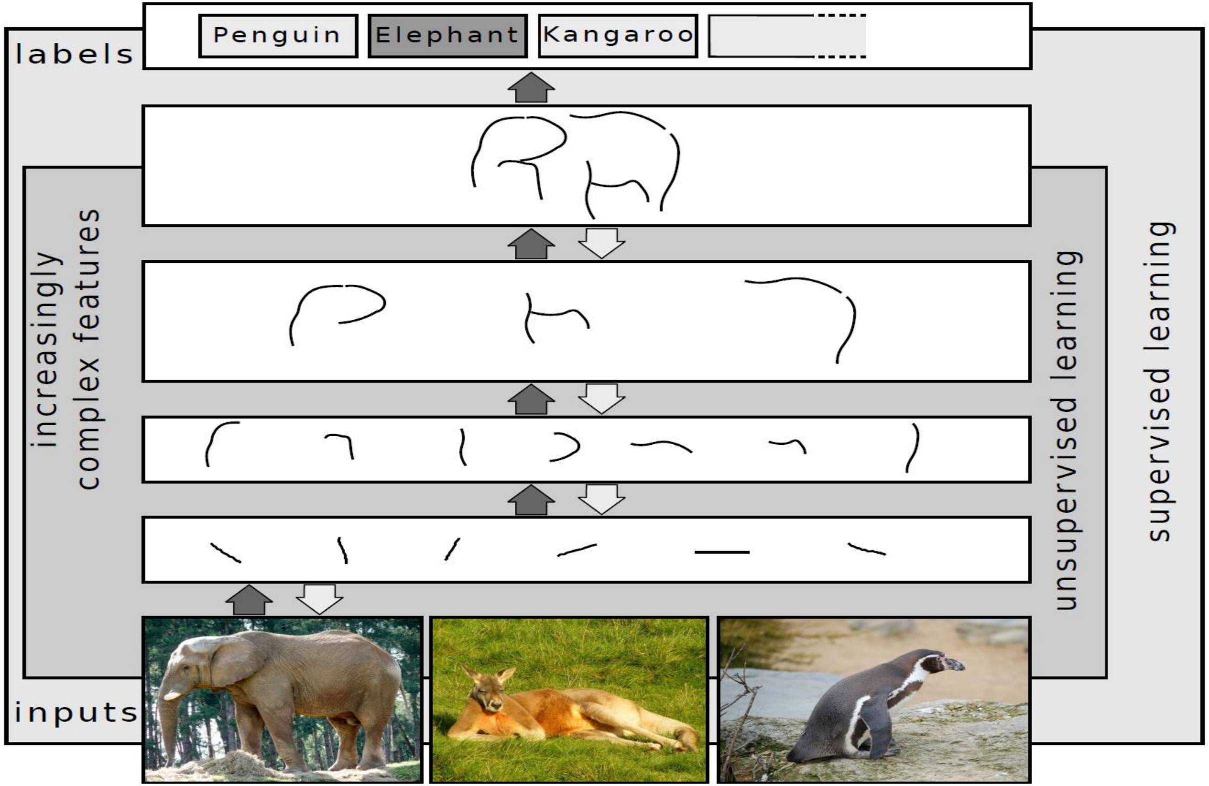
CRF prediction

Ground truth

universitätbonn ais

Deep Learning

- Learning layered represen-tations

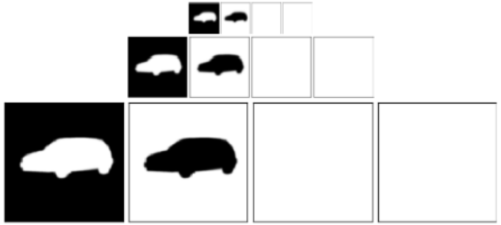


[Schulz; Behnke, KI 2012]

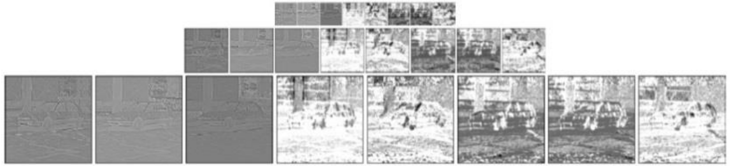
Object-class Segmentation

[Schulz, Behnke, ESANN 2012]

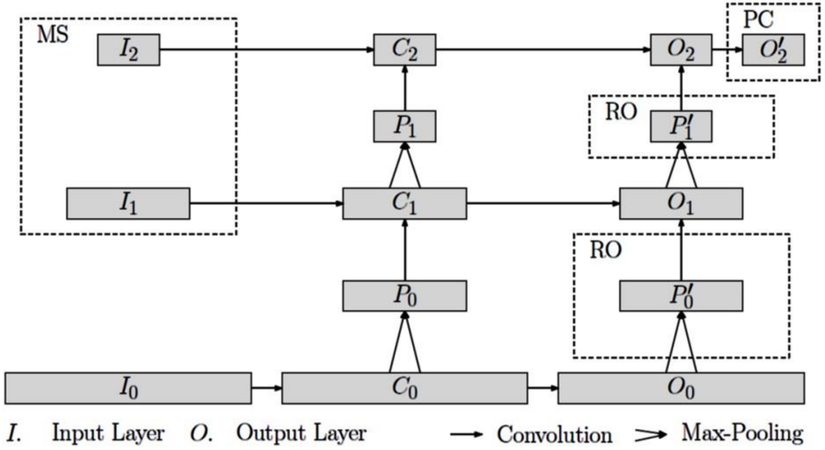
- Class annotation per pixel

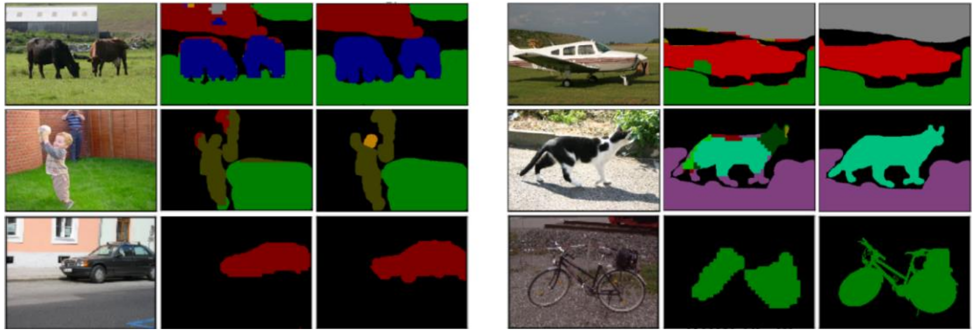


- Multi-scale input channels



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





Input Output Truth

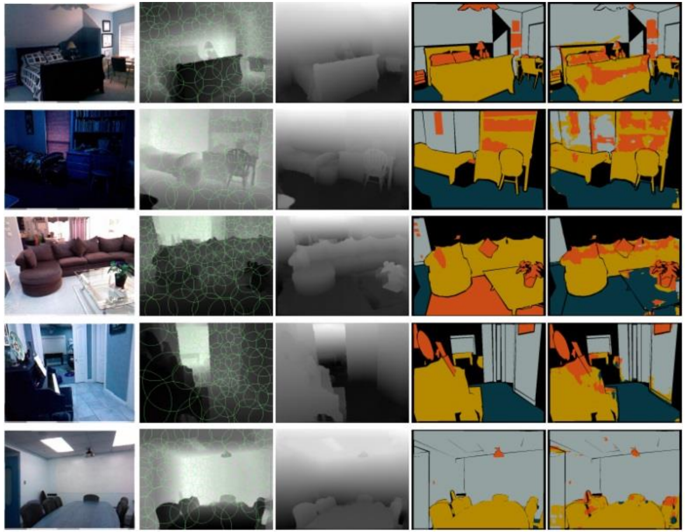
Object Detection in Natural Images

- 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



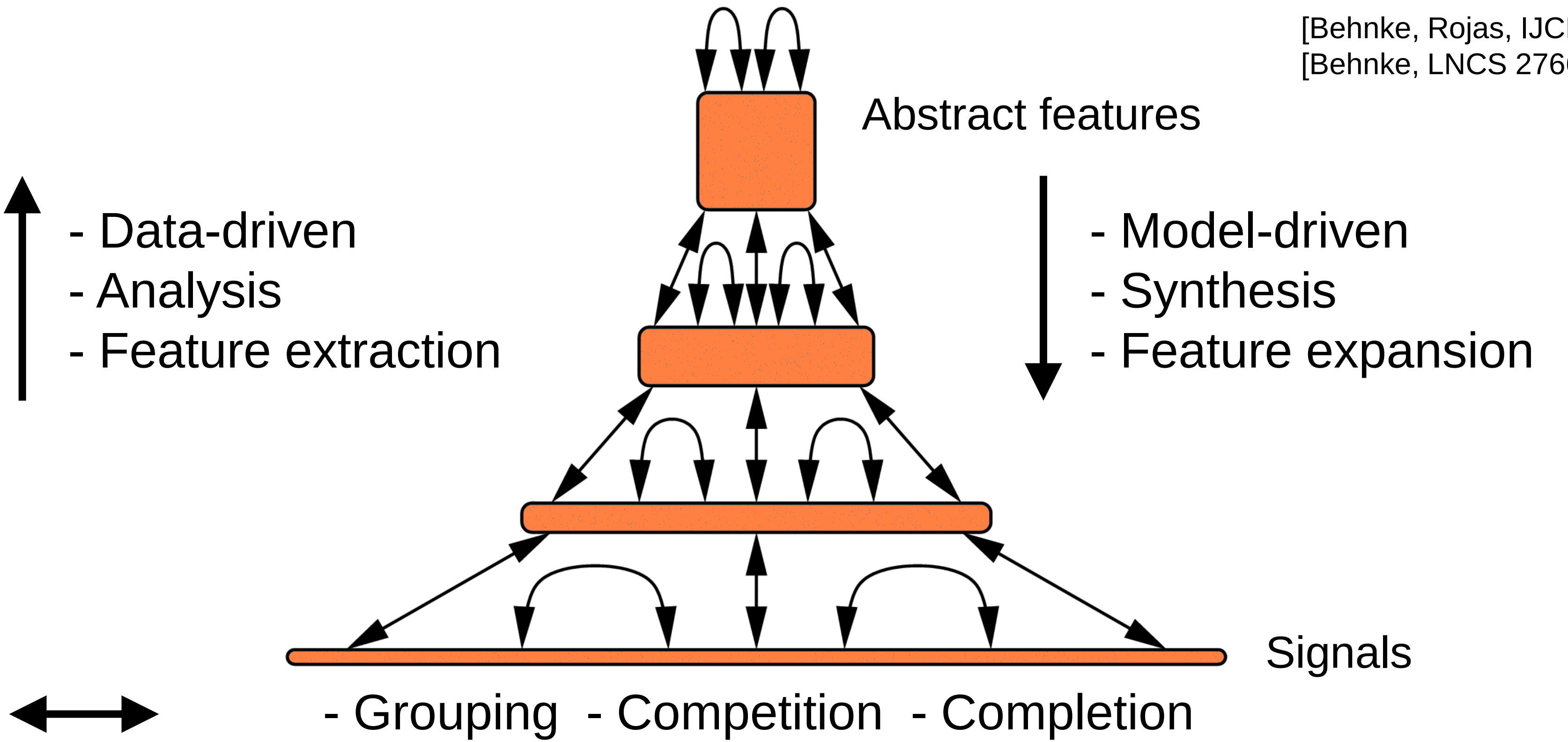
Method	floor	struct	furnit	prop	Class Avg.	Pixel Acc.
CW	84.6	70.3	58.7	52.9	66.6	65.4
CW+DN	87.7	70.8	57.0	53.6	67.3	65.5
CW+H	78.4	74.5	55.6	62.7	67.8	66.5
CW+DN+H	93.7	72.5	61.7	55.5	70.9	70.5
CW+DN+H+SP	91.8	74.1	59.4	63.4	72.2	71.9
CW+DN+H+CRF	93.5	80.2	66.4	54.9	**73.7**	**73.4**
Müller et al.[8]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

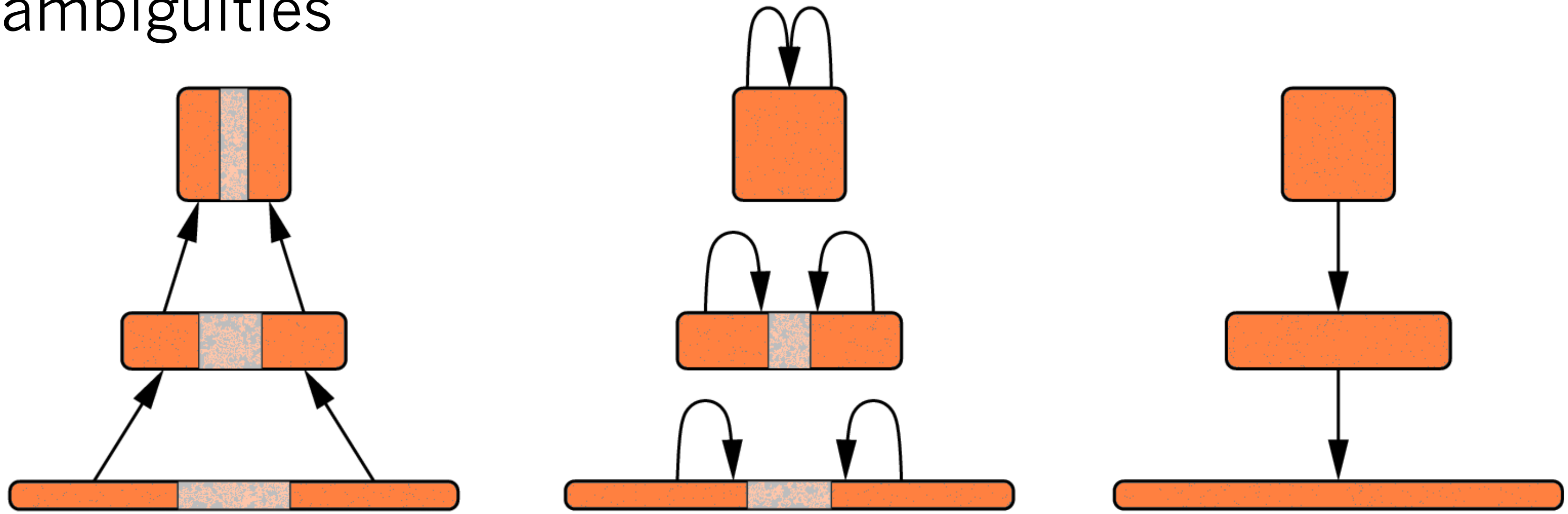
Neural Abstraction Pyramid

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



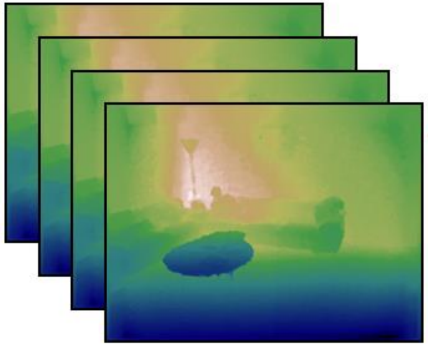
Iterative Image Interpretation

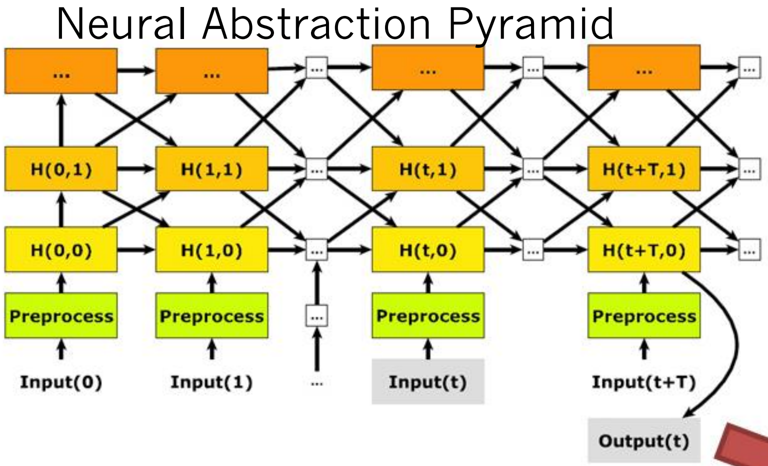
- Interpret most obvious parts first
- Use partial interpretation as context to resolve local ambiguities

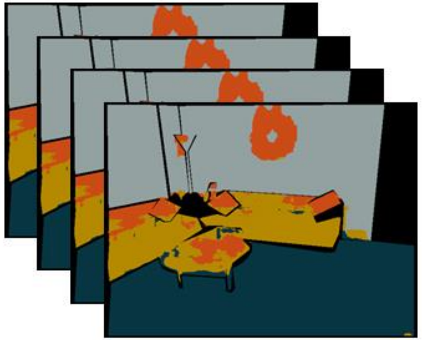


Neural Abstraction Pyramid for RGB-D Video Object-class Segmentation

- Recursive computation is efficient for temporal integration



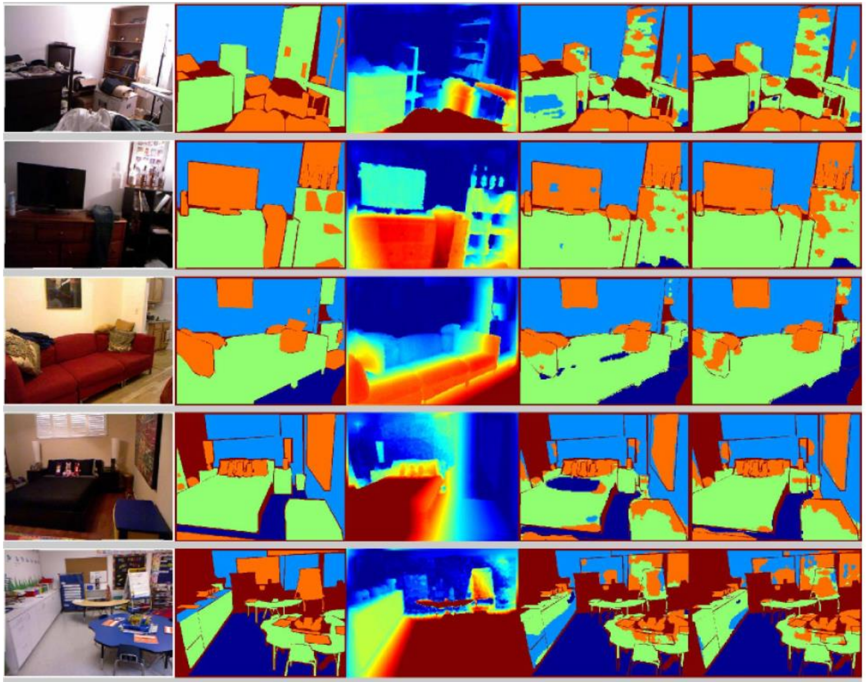




[Pavel, Schulz, Behnke, Neural Networks 2017]

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

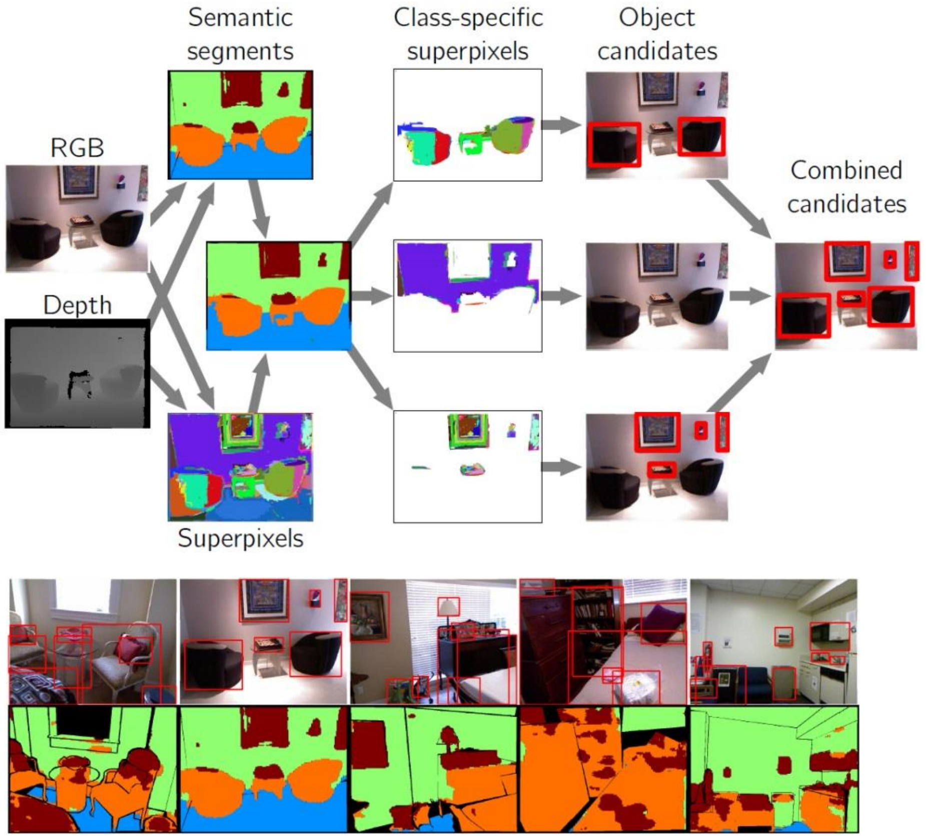
- New **geometric** feature: distance from wall
- **Semantic** features pretrained from ImageNet
- Both help significantly



[Husain et al. RA-L 2016]

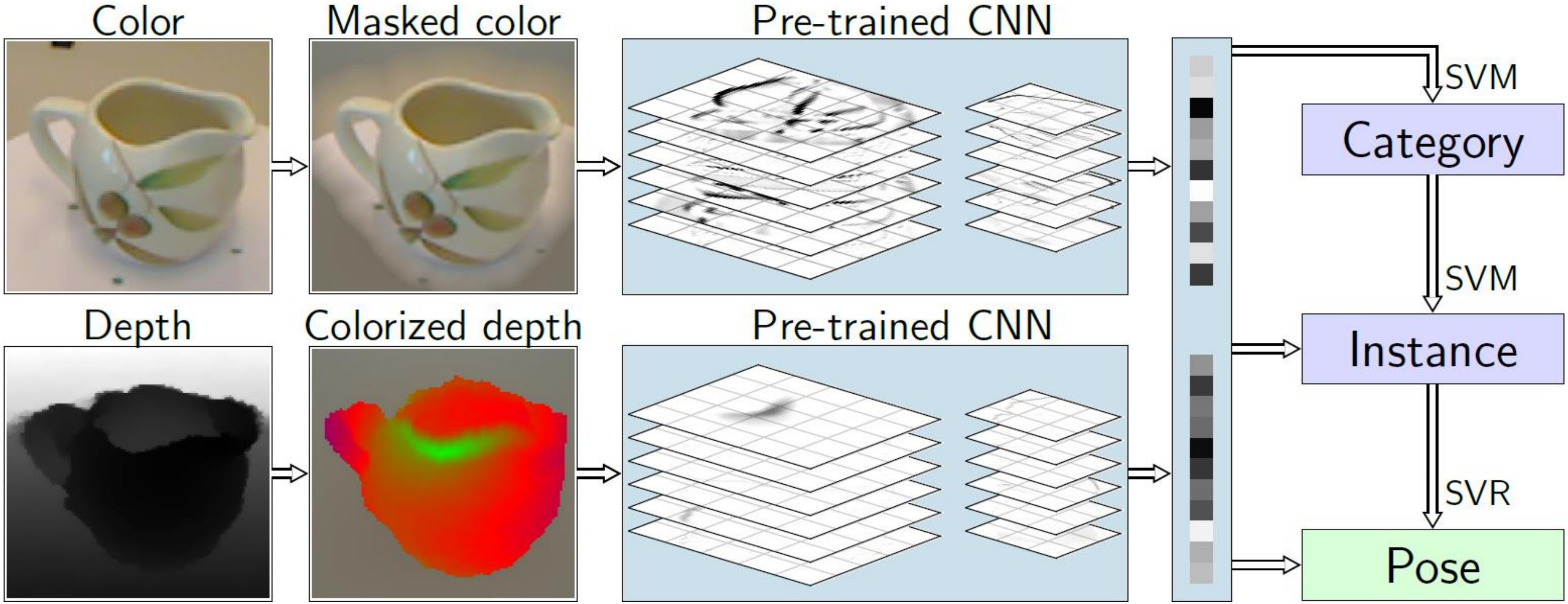
Semantic Segmentation Priors for Object Discovery

- 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. ICPR 2016]

RGB-D Object Recognition and Pose Estimation

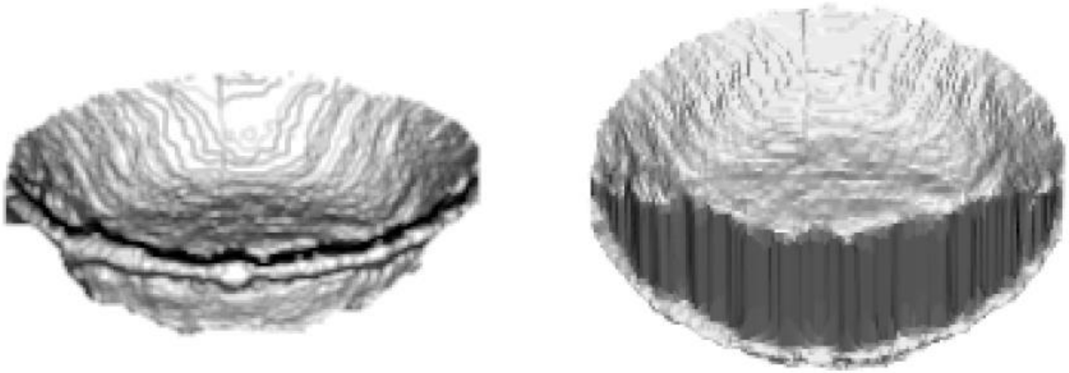


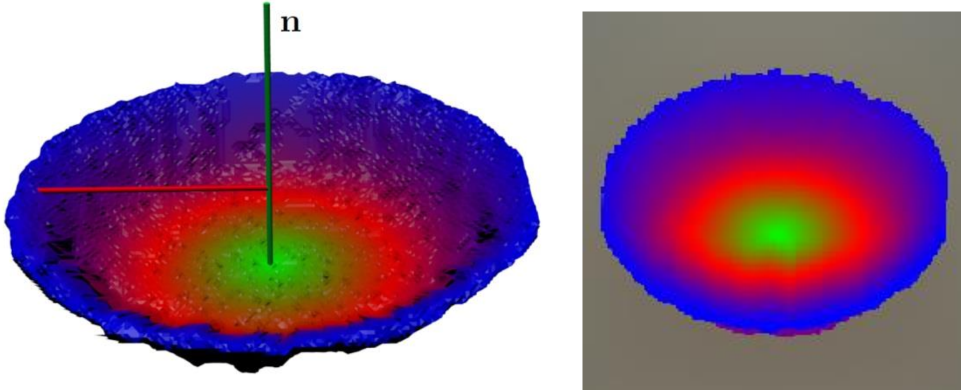
[Schwarz, Schulz, Behnke, ICRA2015]

Canonical View, Colorization

- Objects viewed from different elevation
- Render canonical view

- Colorization based on distance from center vertical

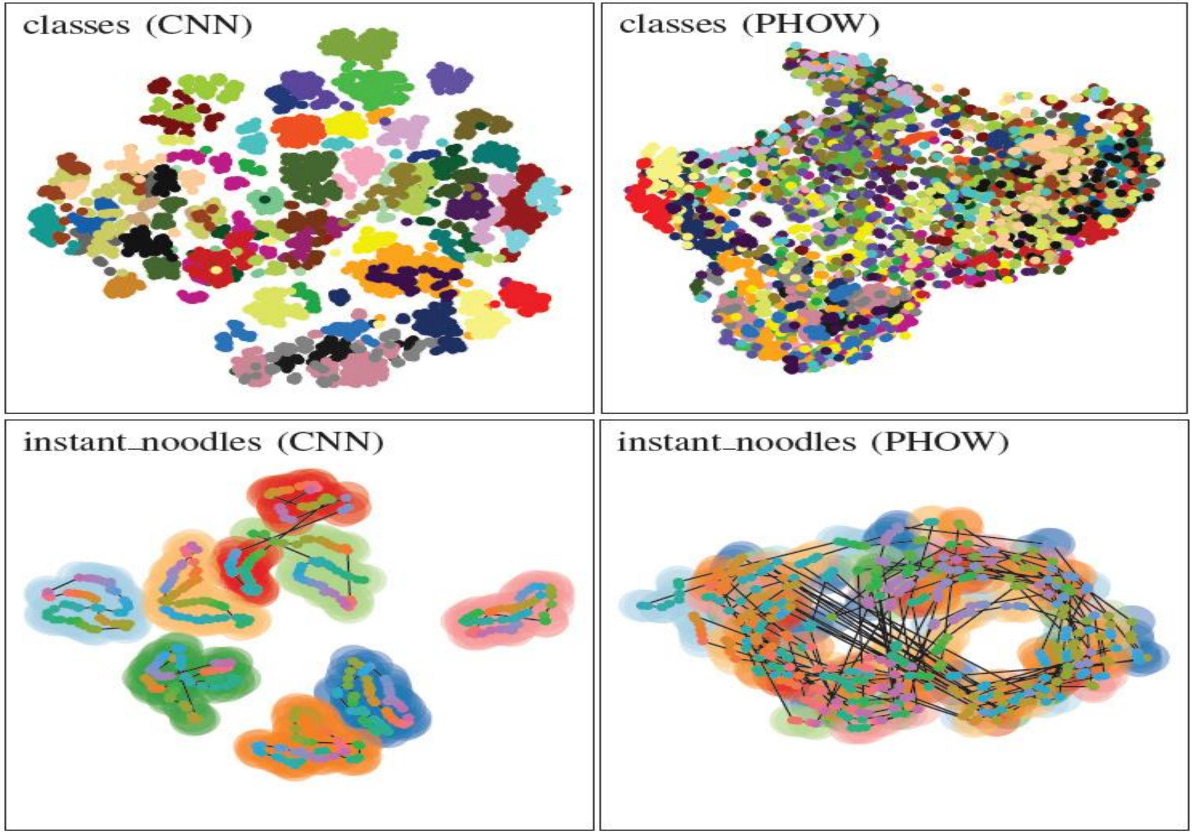




[Schwarz, Schulz, Behnke, ICRA2015]

Pretrained Features Disentangle Data

- t-SNE embedding



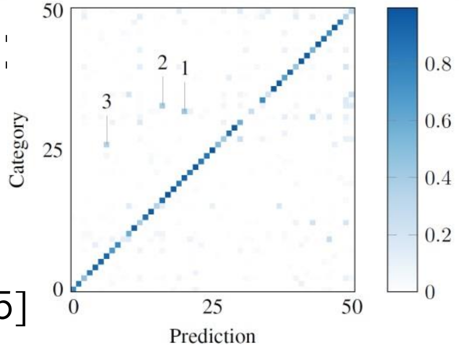
[Schwarz, Schulz, Behnke ICRA2015]

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:



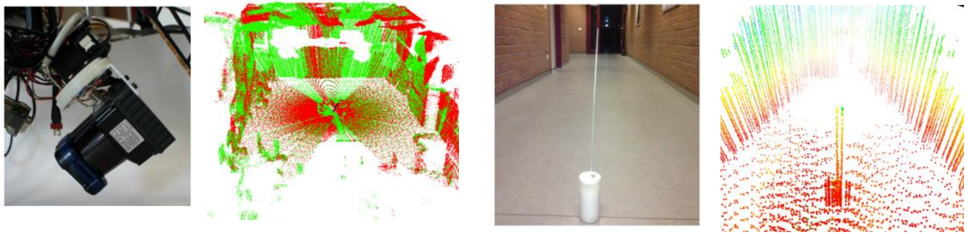
1: pitcher / coffe mug

2: peach / sponge

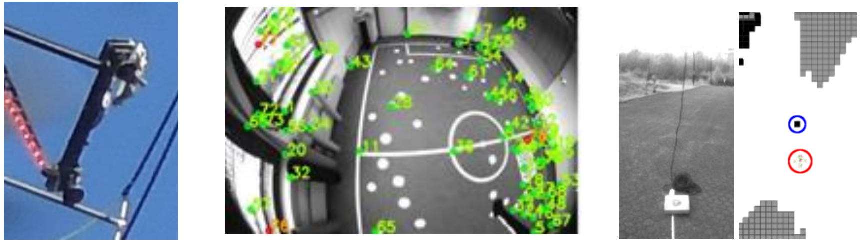
[Schwarz, Schulz, Behnke, ICRA2015]

Autonomous Flight Near Obstacles

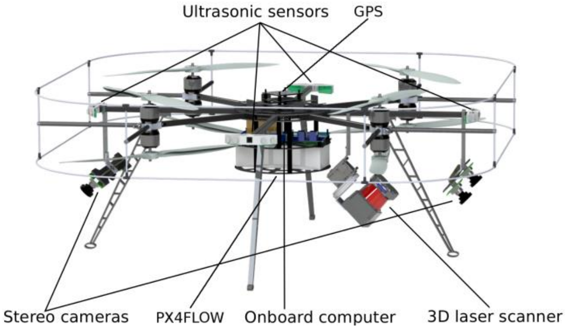
- Multimodal obstacle detection
  - 3D laser scanner

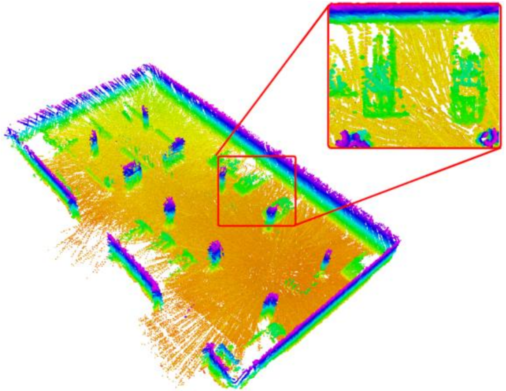


  - Stereo cameras



  - Ultrasound

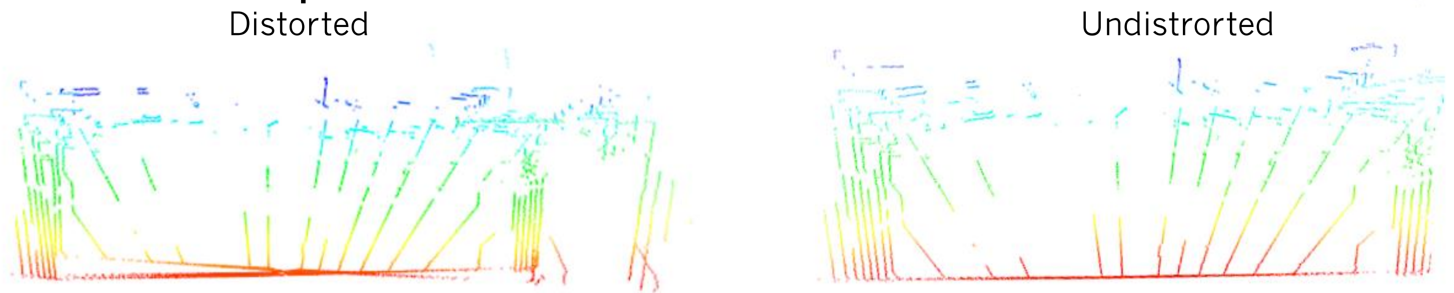




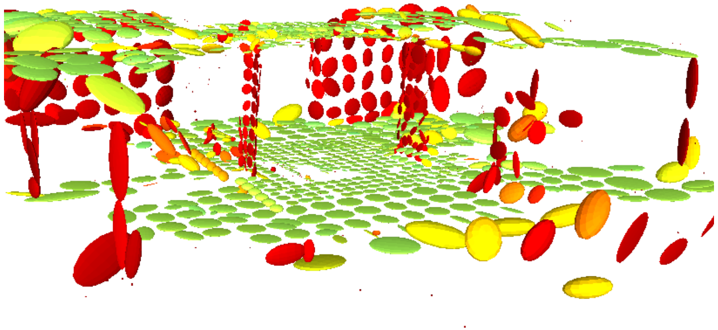
[Droeschel et al.: Journal of Field Robotics, 2015]

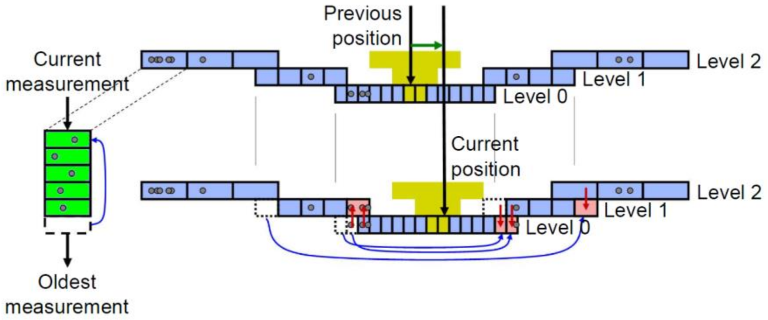
Egocentric Laser-based 3D Mapping

- Motion compensation



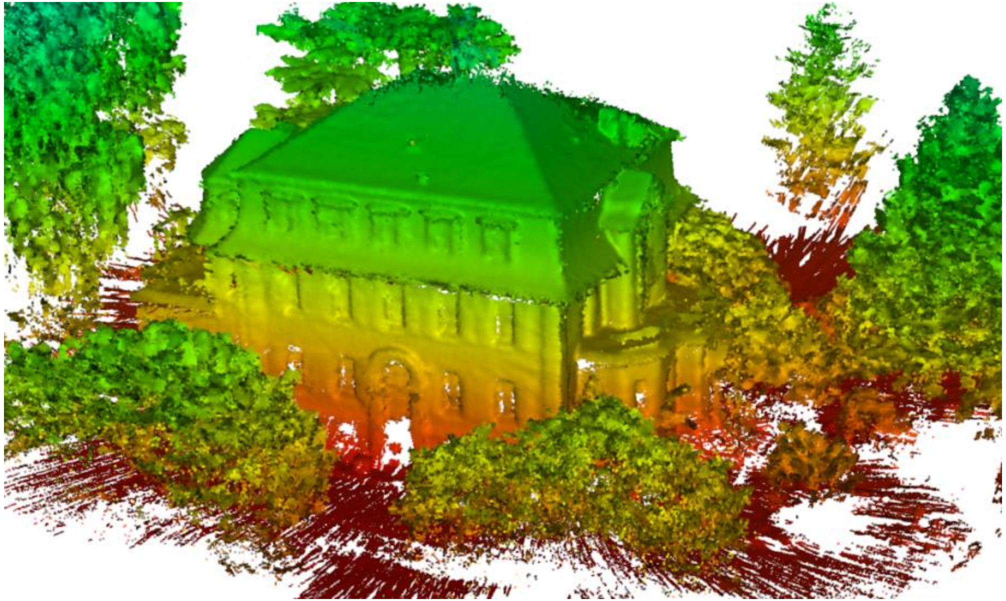
- Local multiresolution surfel maps





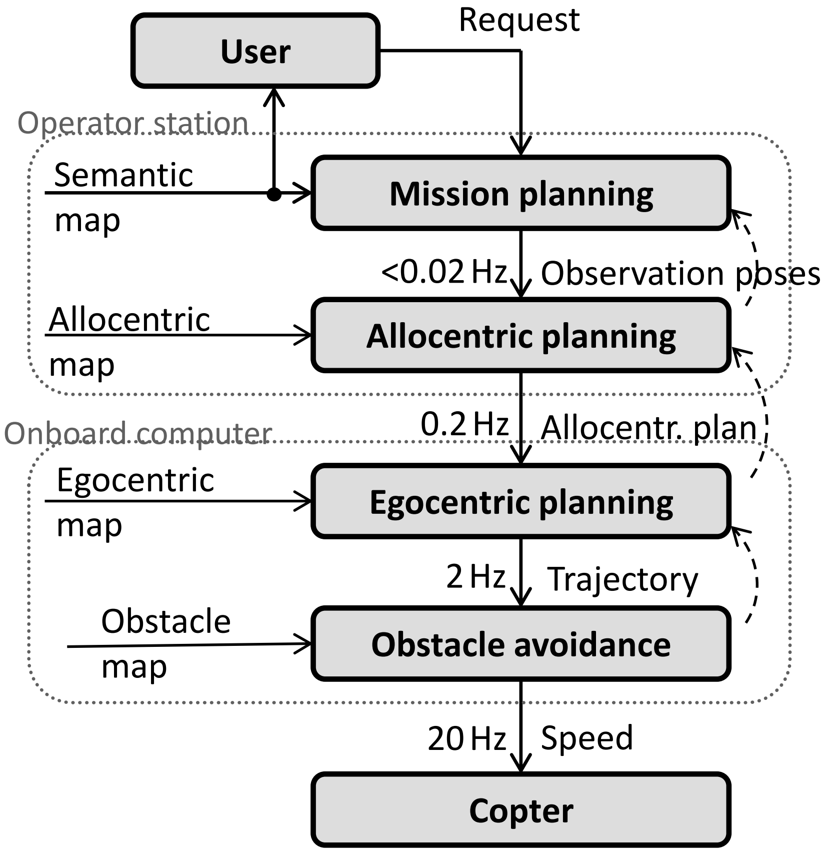
Allocentric 3D Map

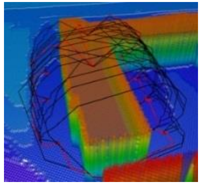
- Registration of egocentric maps
- Global optimization of registration error by GraphSLAM



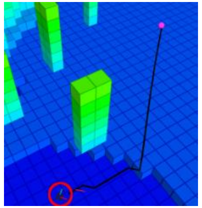
[Droeschel et al. JFR 2016]

Hierarchical Navigation

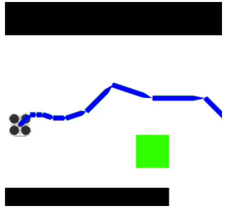




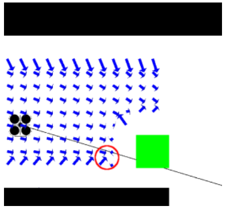
Mission plan



Allocentric planning



Egocentric planning

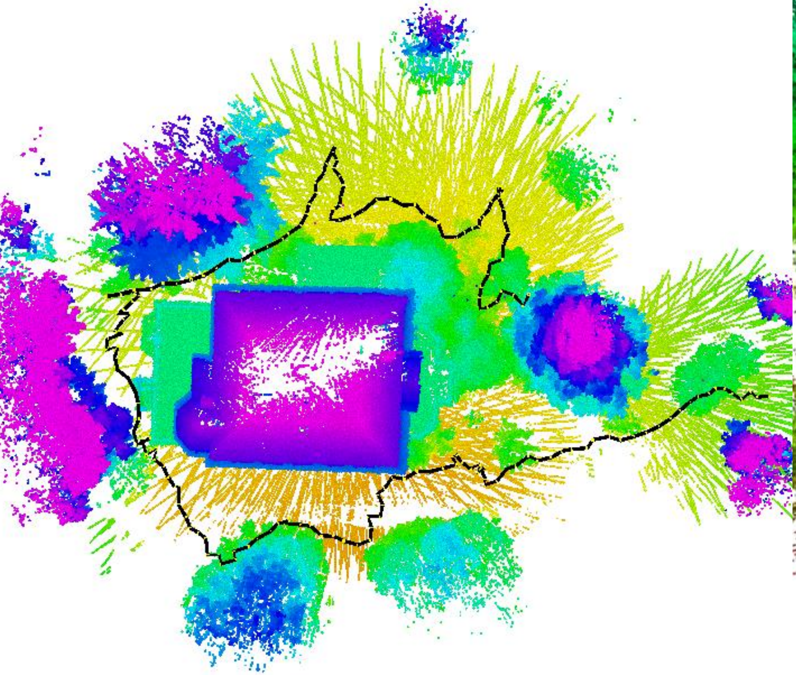


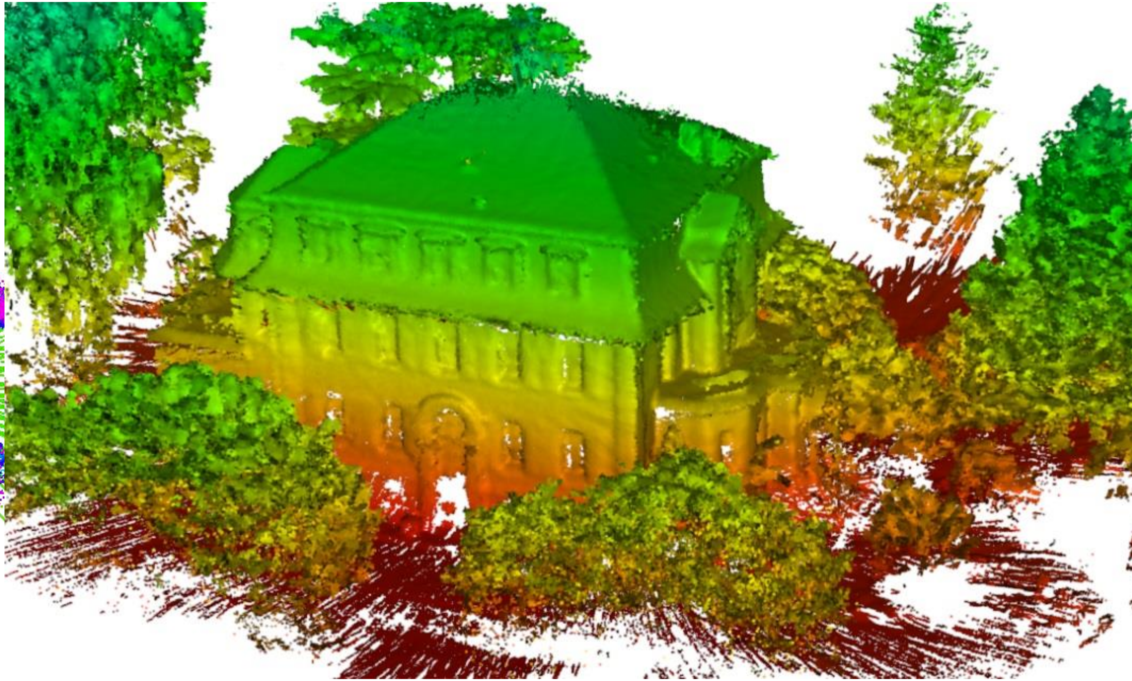
Obstacle avoidance

Mapping on Demand

Autonomous Flight to Planned View Poses

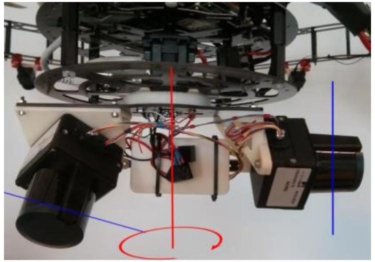
3D Simultaneous Localization and Mapping

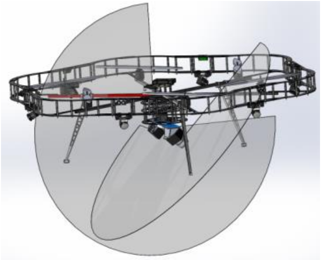


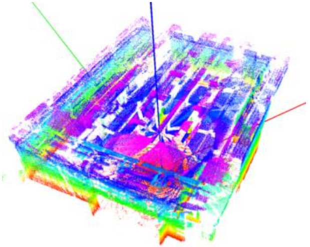


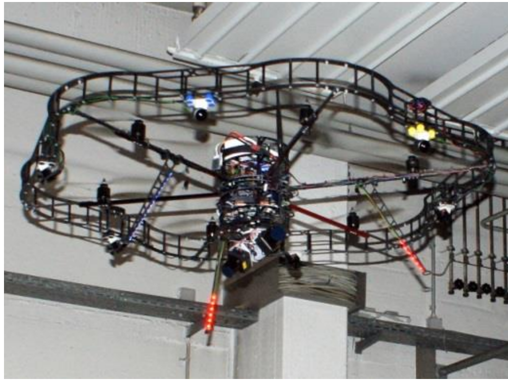
Autonomous Flight in Warehouses

- Dual 3D laser scanner

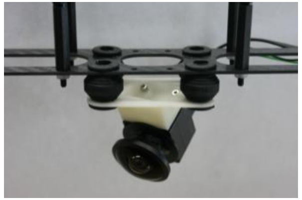


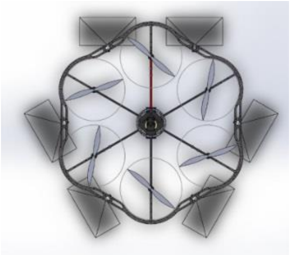


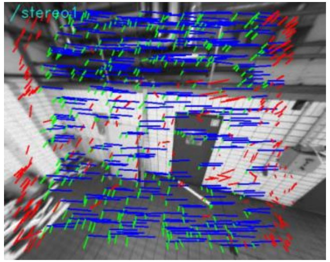


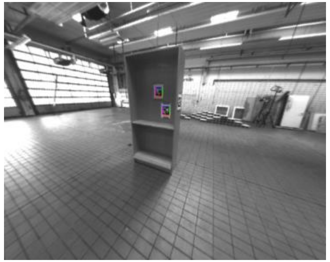


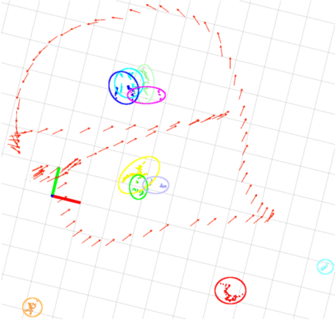
- Omnidirectional cameras





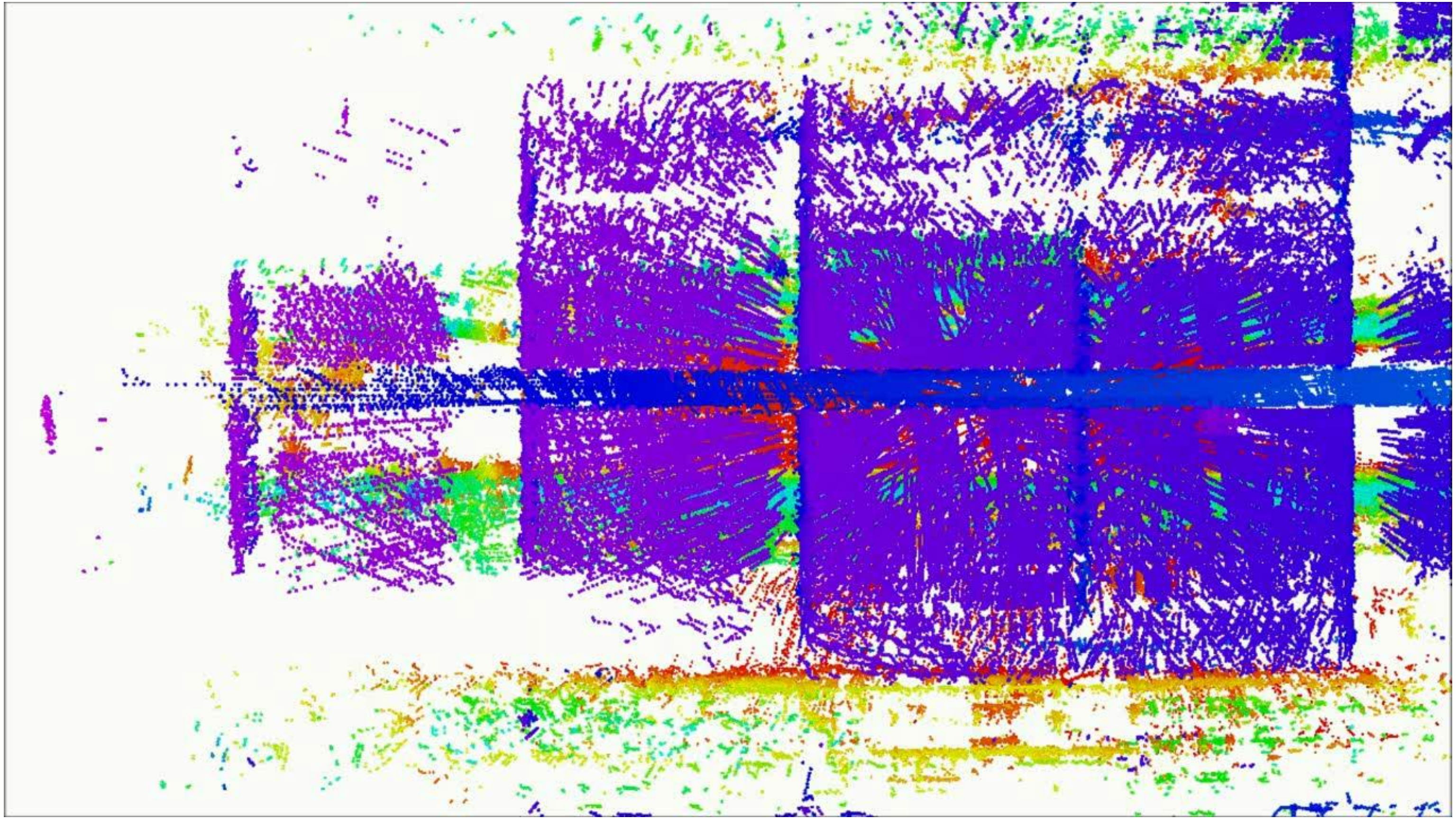




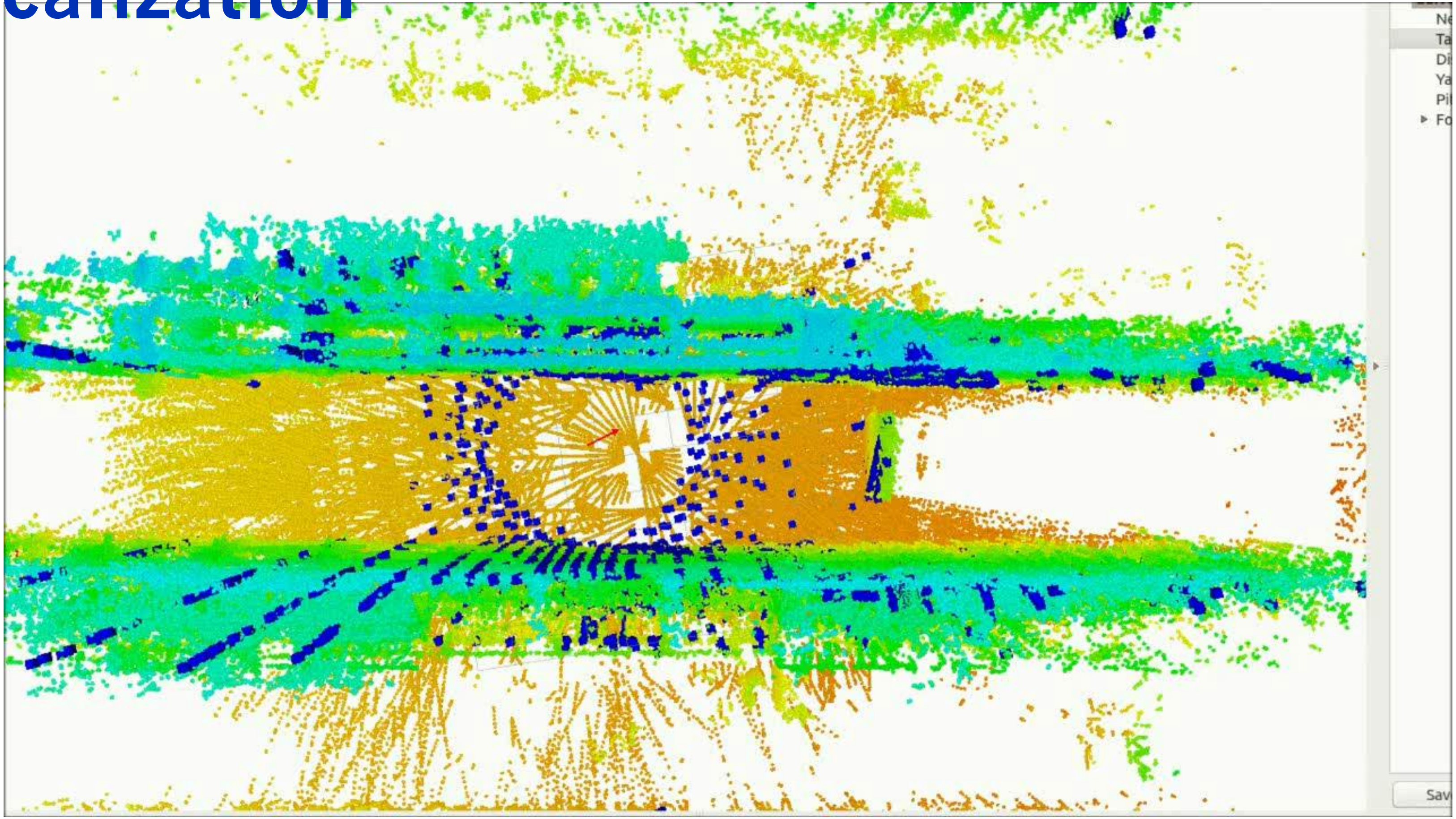


- RFID reader

3D Map



Localization

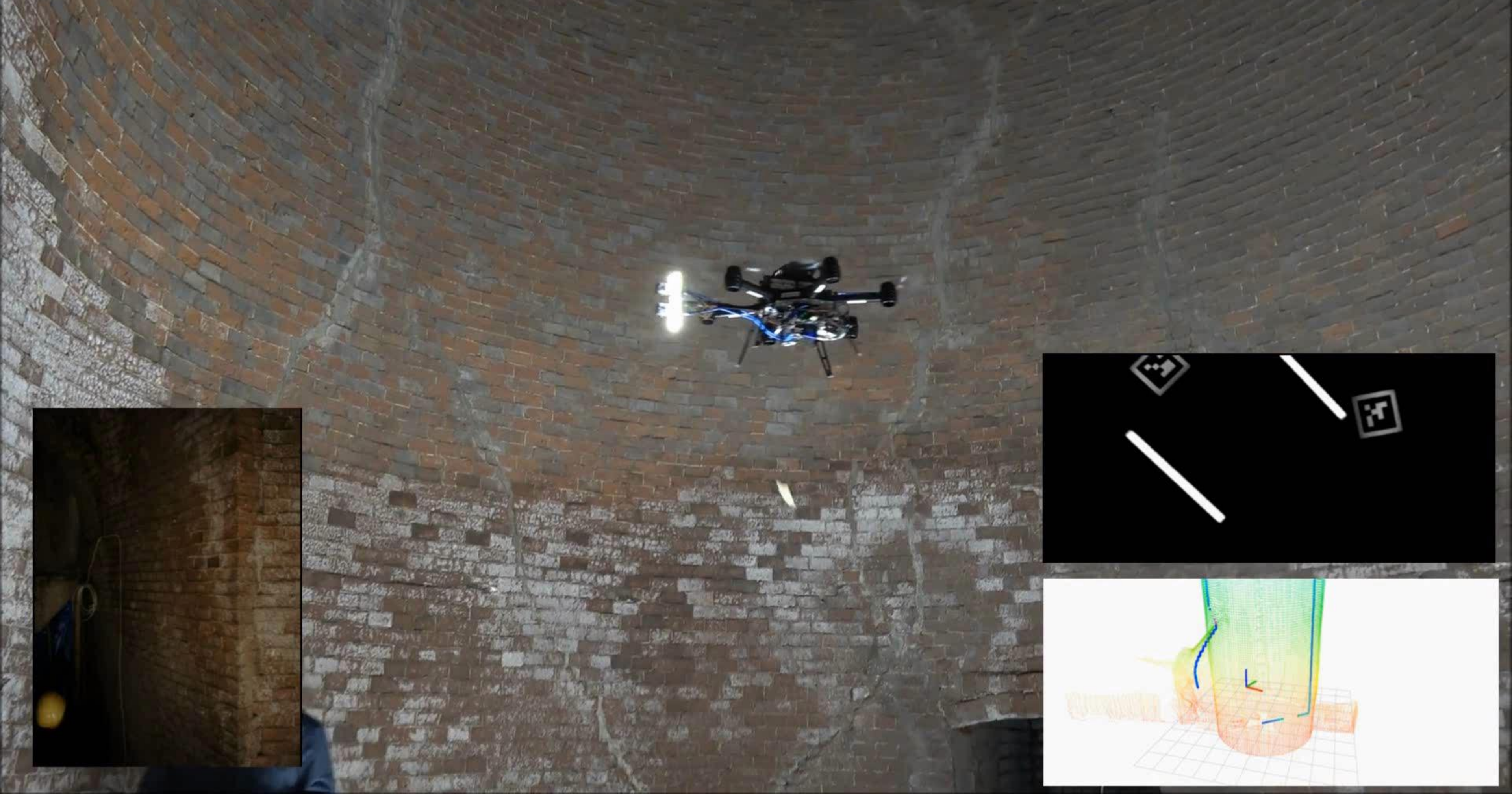



Autonomous Mission in Warehouse

DJI Matrice 600 with Velodyne Puck

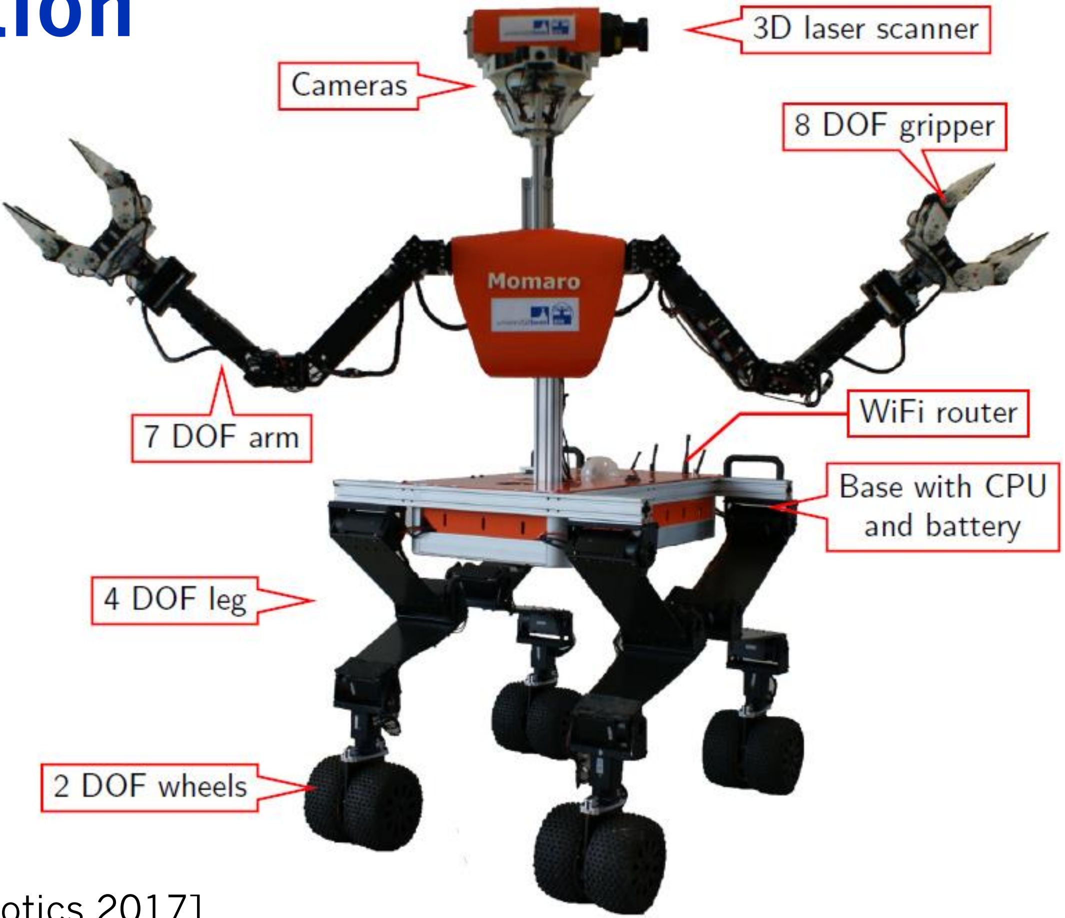
InventAIRy Final Demonstration

EuRoC Challenge 3: ChimneySpector



Mobile Manipulation Robot Momaro

- Four compliant legs ending in pairs of steerable wheels
- Anthropomorphic upper body
- Sensor head
  - 3D laser scanner
  - IMU, cameras



[Schwarz et al. Journal of Field Robotics 2017]

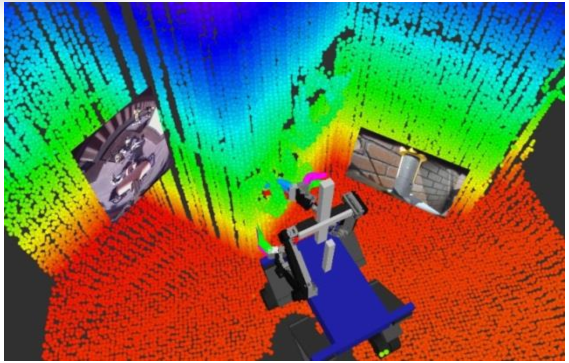
Driving a Vehicle

Egress

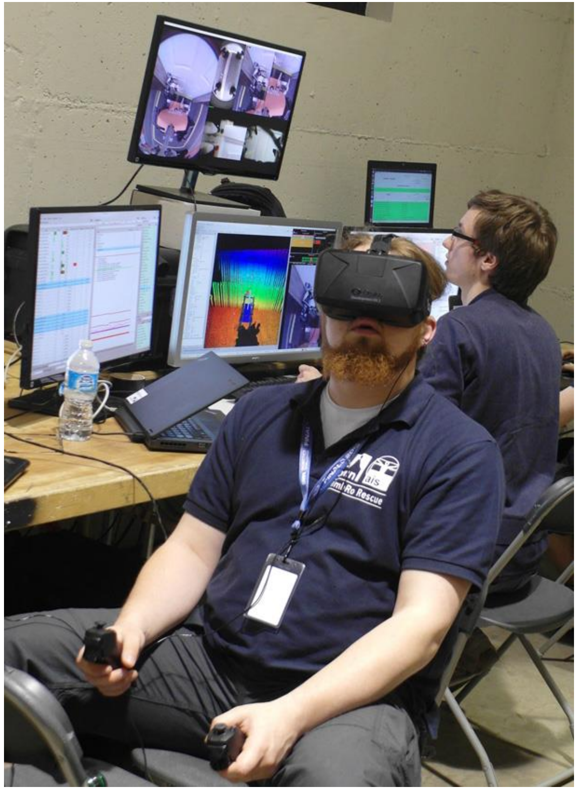
Manipulation Operator Interface

- 3D head-mounted display

- 3D environment model + images



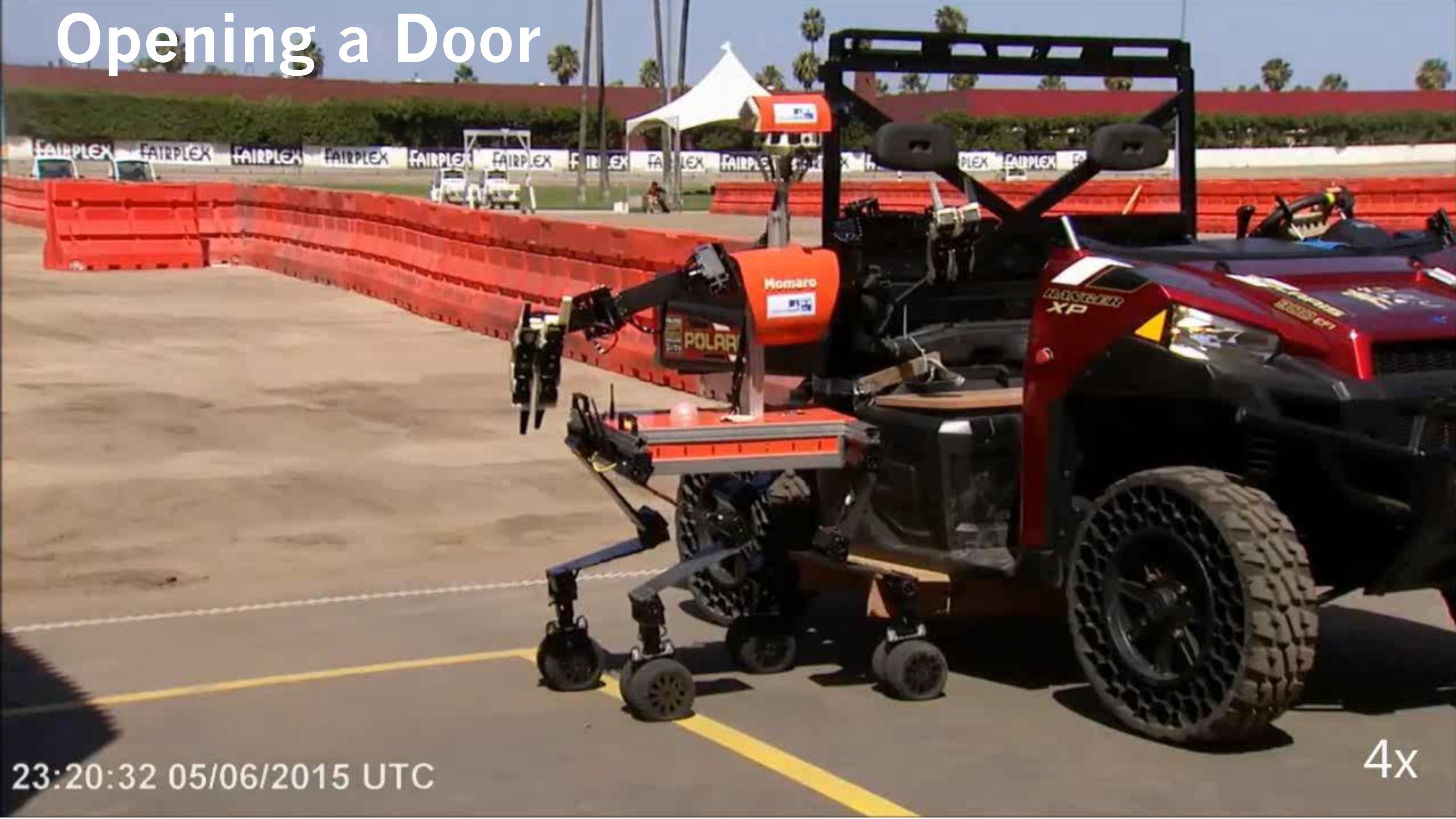
- 6D magnetic tracker



[Rodehutskors et al., Humanoids 2015]

universitätbonn ais

Opening a Door

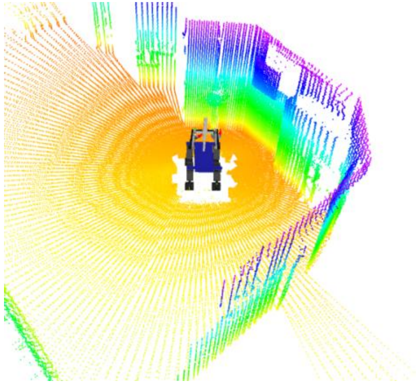


Local Multiresolution Surfel Map

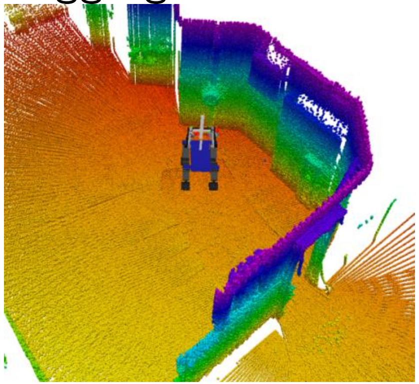
- Registration and aggregation of 3D laser scans
- Local multi-resolution grid
- Surfel in grid cells

[Droeschel et al., Robotics and Autonomous Systems 2017]

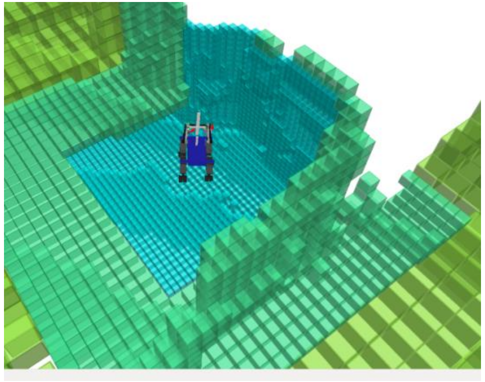
3D scan



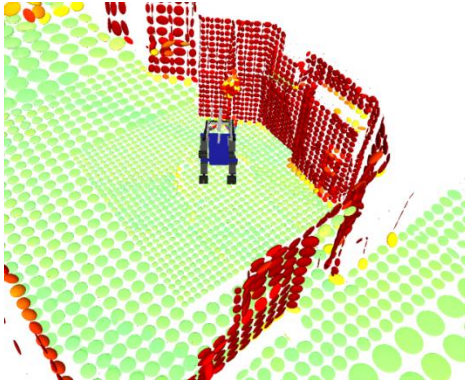
Aggregated scans



Multiresolution grid

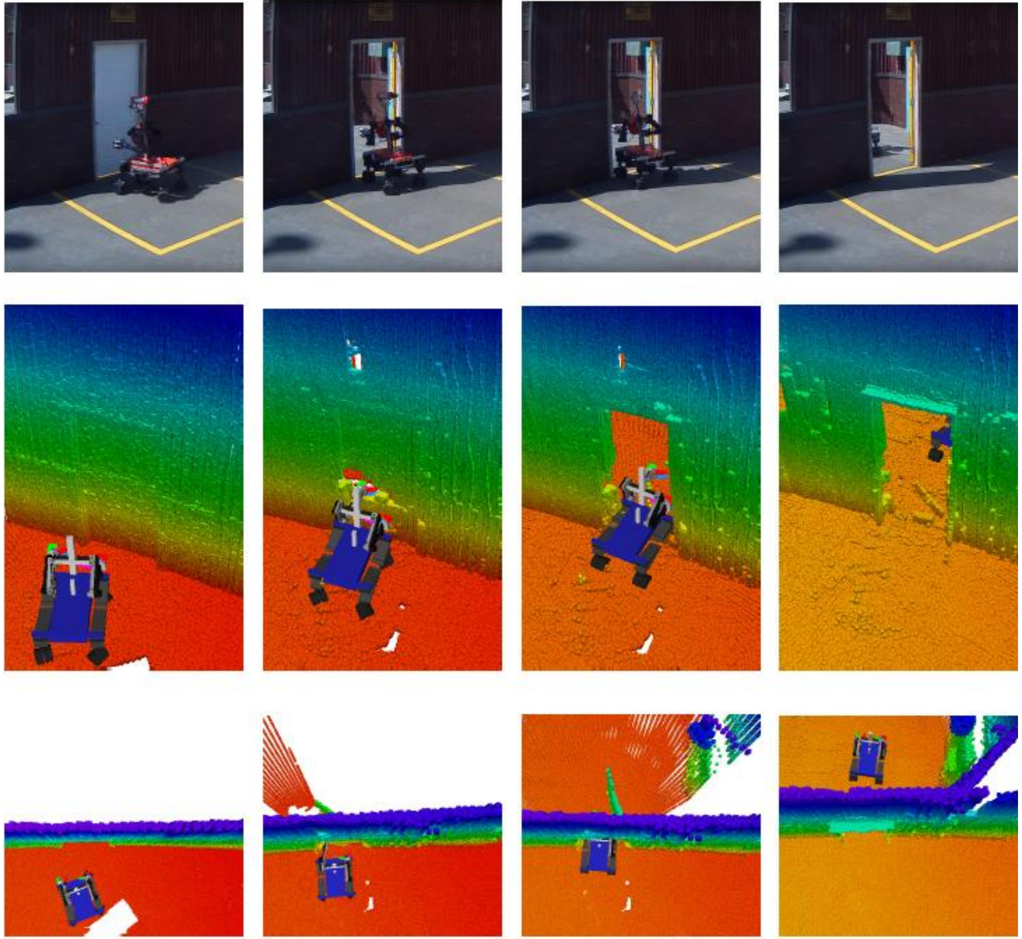


Surfels



Filtering Dynamic Objects

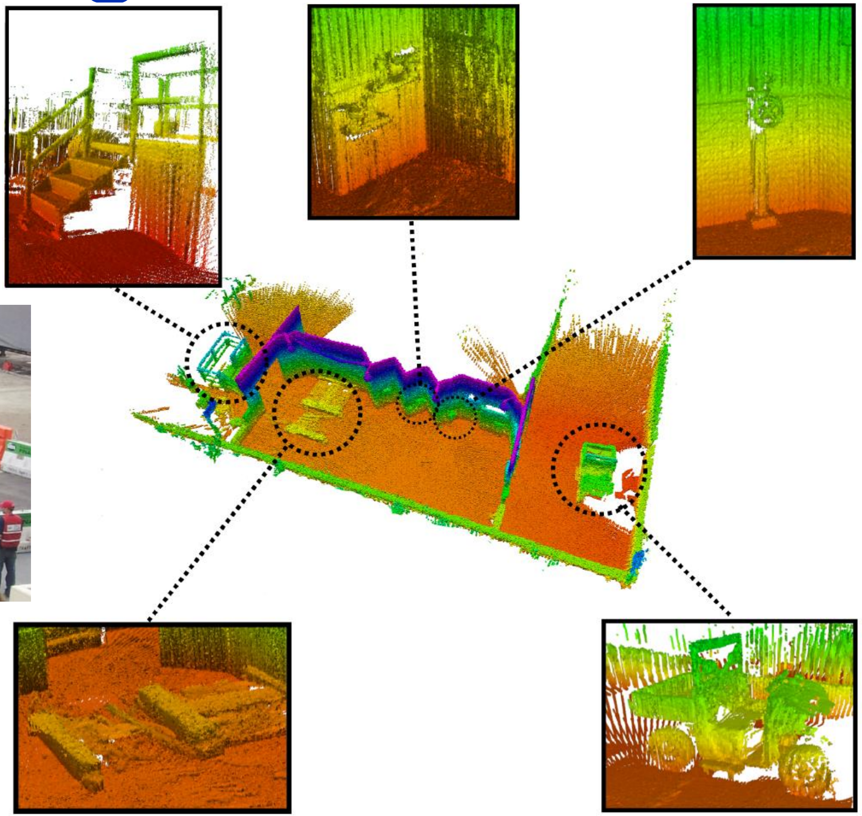
- Maintain occupancy in each cell


1 scan (5 s) 2 scans (10 s) 5 scans (25 s)

universitätbonn ais

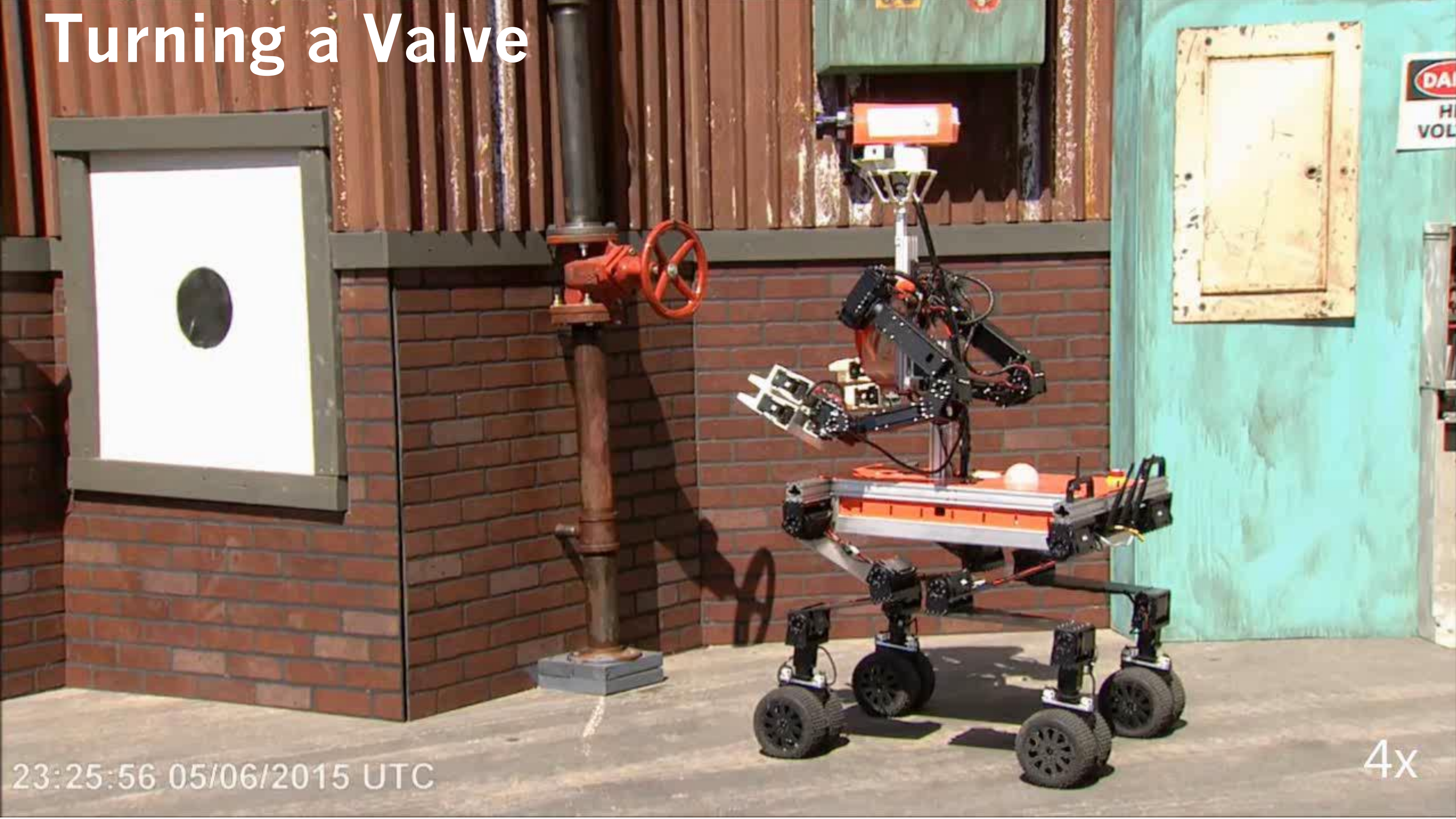
Allocentric 3D Mapping

- Registration of egocentric maps by graph optimization



[Droeschel et al., Robotics and Autonomous Systems 2017]

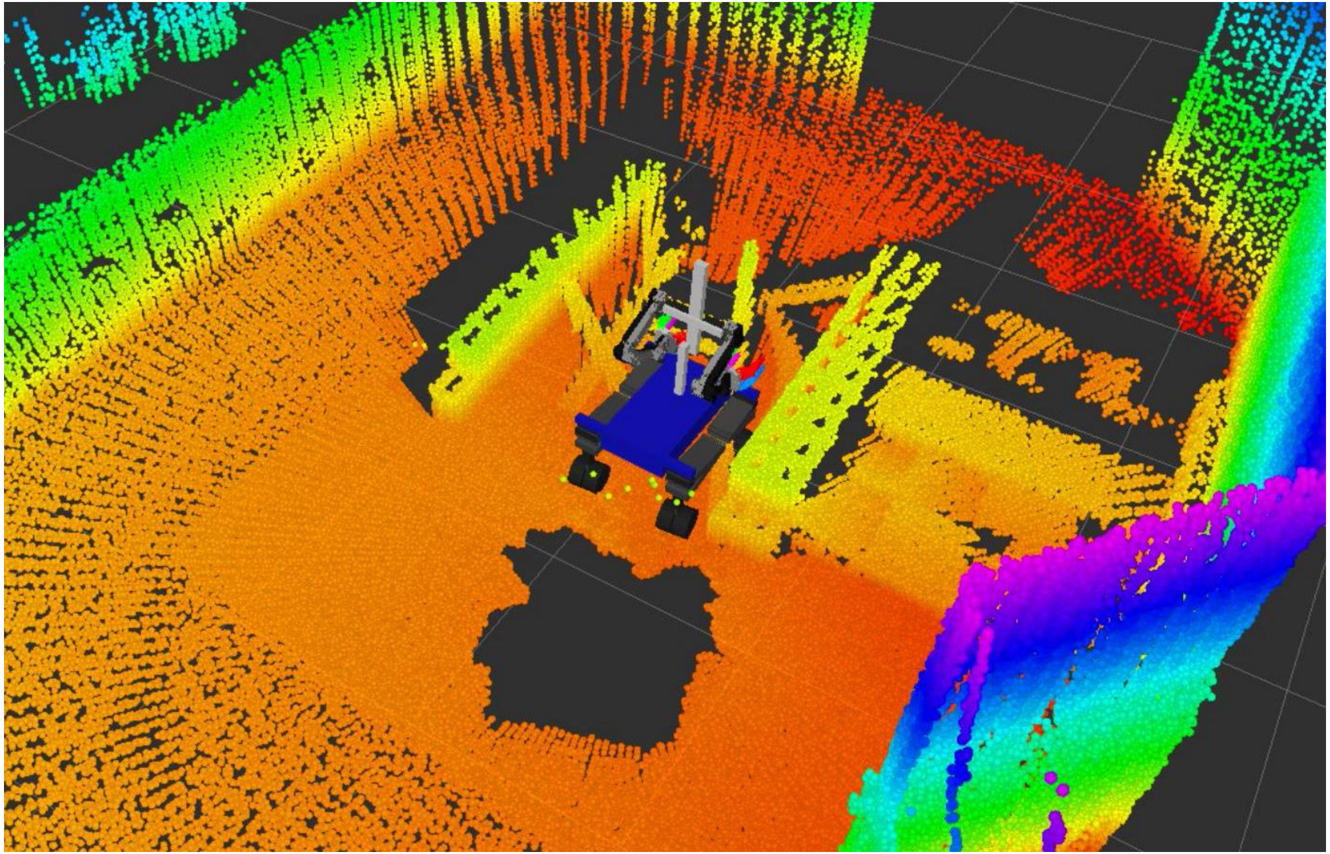
universitätbonn ais


Turning a Valve
23:25:56 05/06/2015 UTC
4x

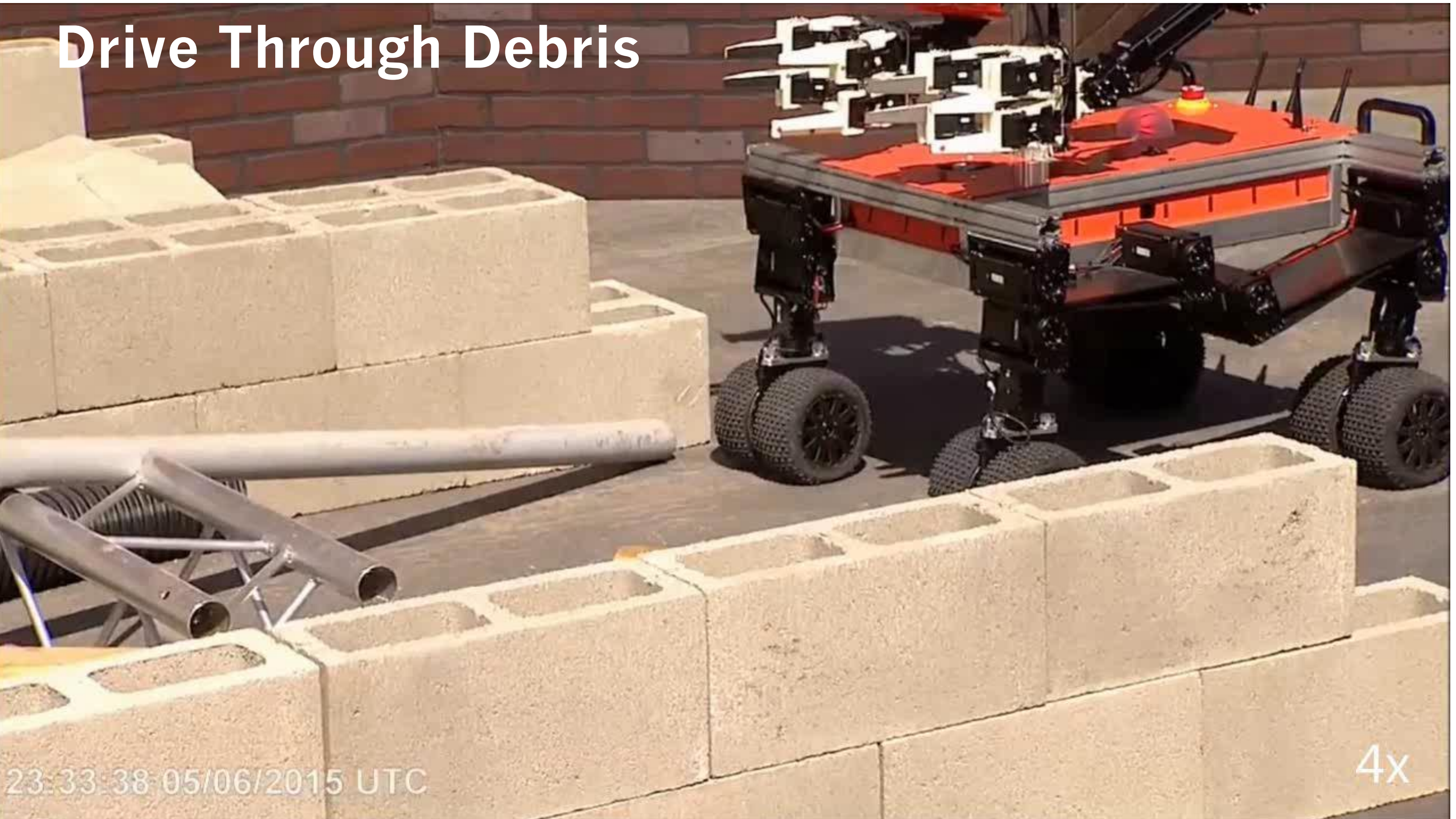
Operating a Switch

Plug Task

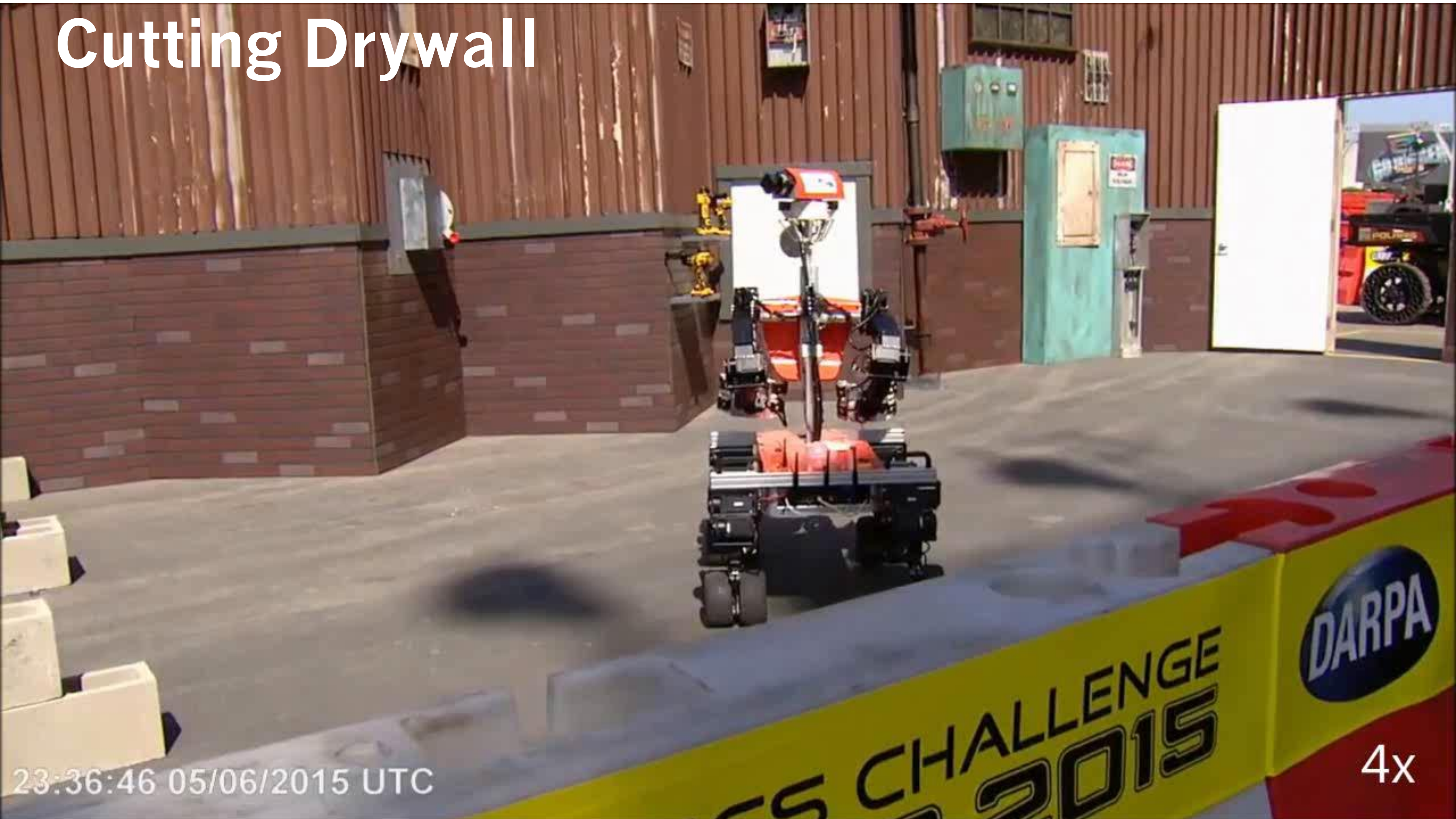
Debris Tasks



Drive Through Debris



Cutting Drywall



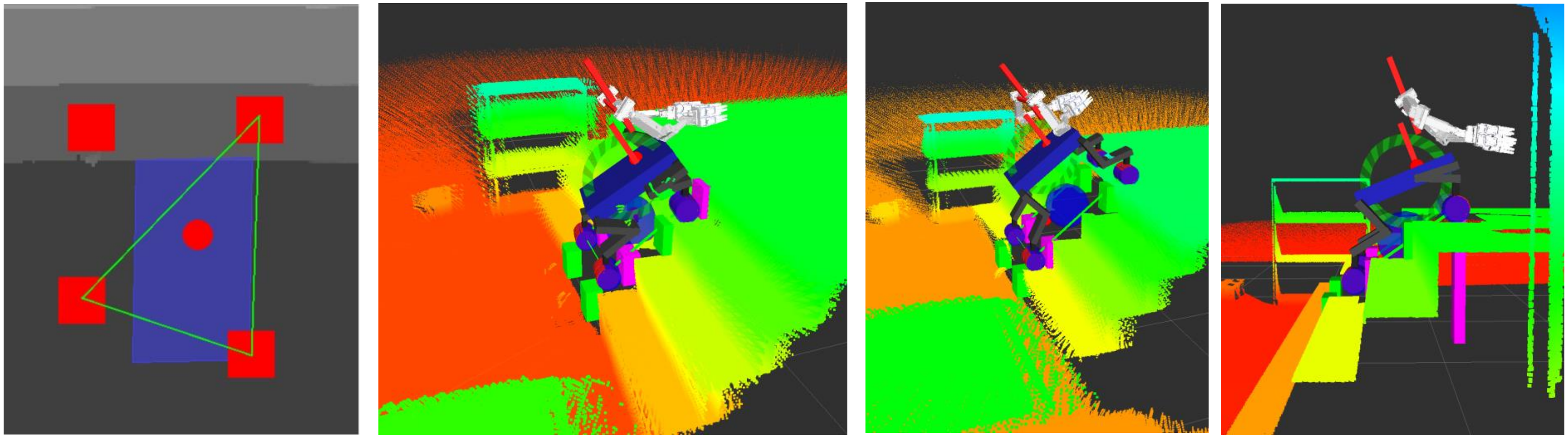
Team NimbRo Rescue

Best European Team (4th place overall), solved seven of eight tasks in 34 minutes

universitätbonn ais

Stair Climbing

- Determine leg that most urgently needs to step
- Weight shift: sagittal, lateral, driving changes support
- Step to first possible foot hold after height change



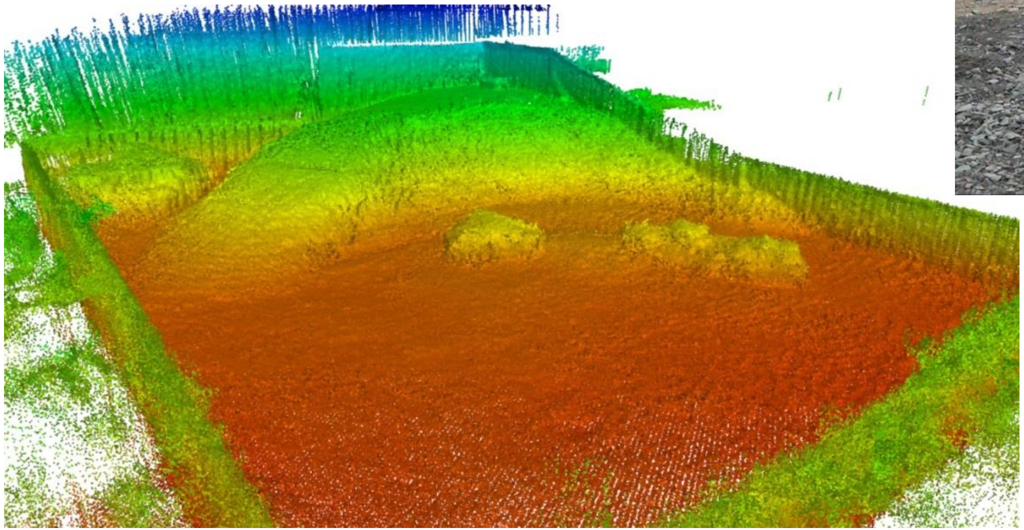
[Schwarz et al., ICRA 2016]

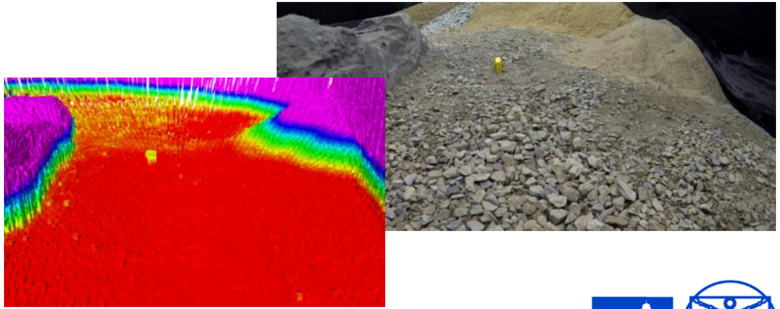
Stair Crawling

DLR SpaceBot Cup 2015

- Mobile manipulation in rough terrain

[Schwarz et al., Frontiers on Robotics and AI 2016]



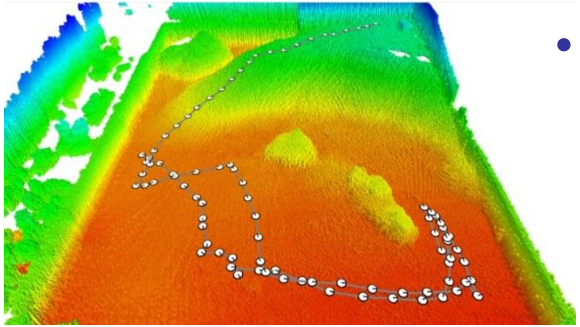


universitätbonn ais

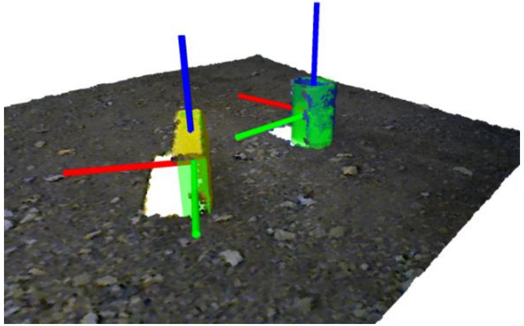
DLR SpaceBot Camp 2015

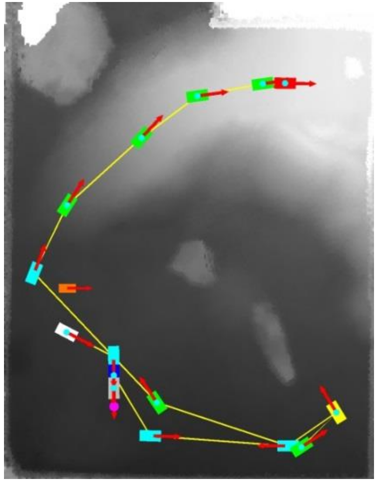
Autonomous Mission Execution

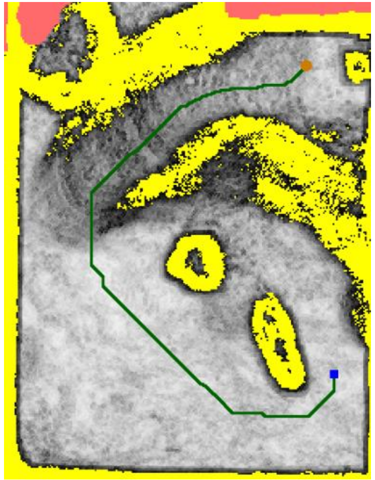
- 3D mapping, localization, mission and navigation planning

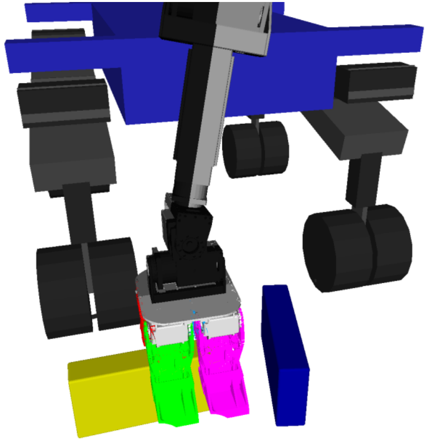


- 3D object perception and grasping







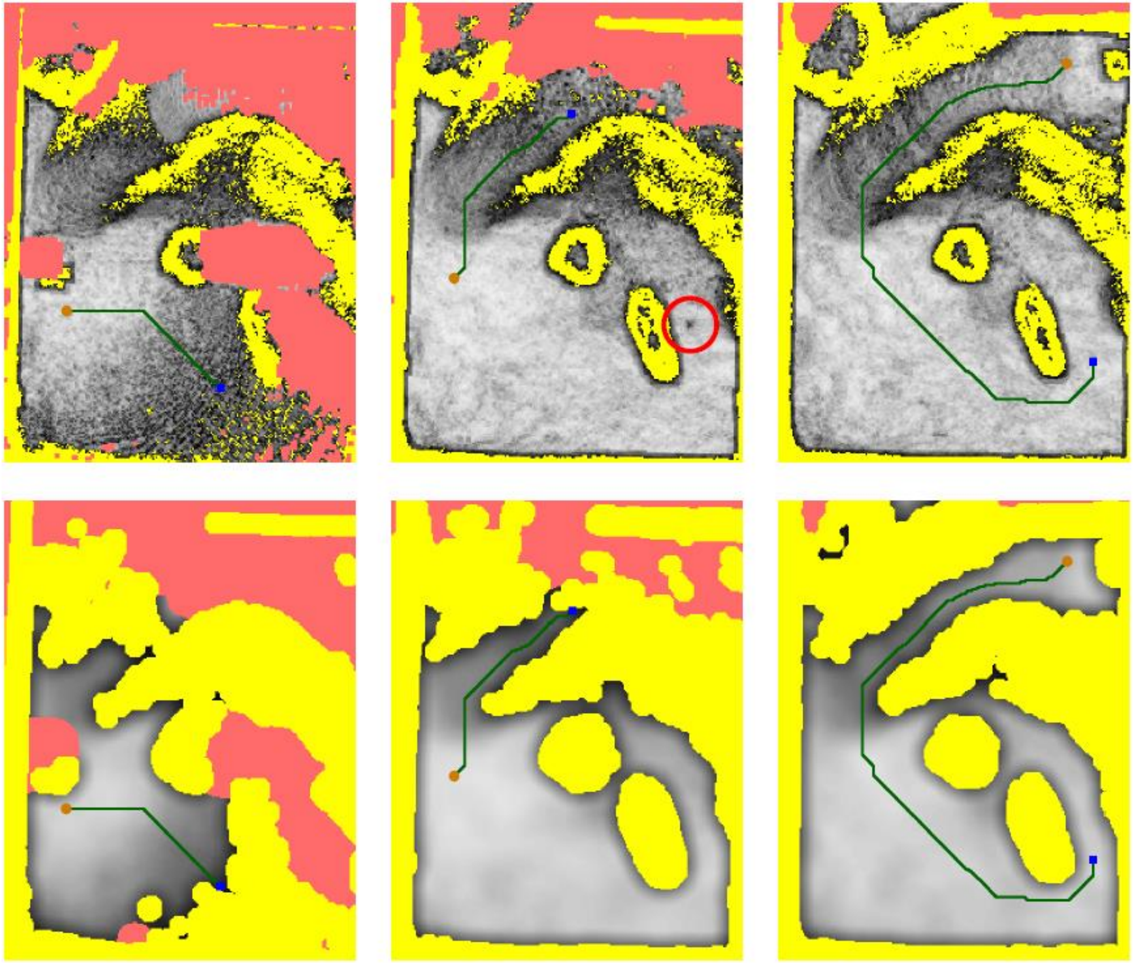


[Schwarz et al. Frontiers 2016]

universitätbonn ais

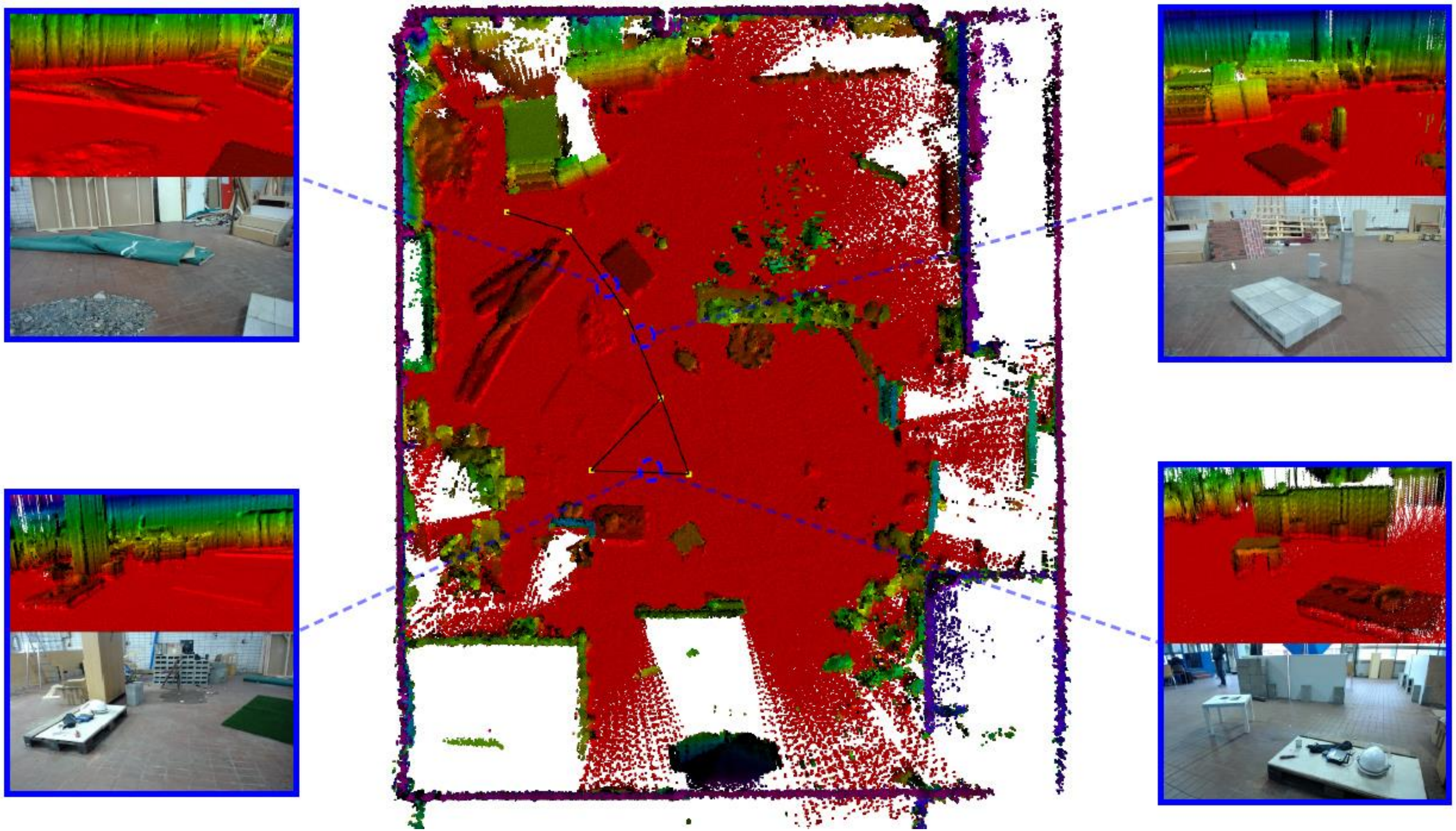
Navigation Planning

- Costs from local height differences
- A* path planning



[Schwarz et al., Frontiers in Robotics and AI 2016]

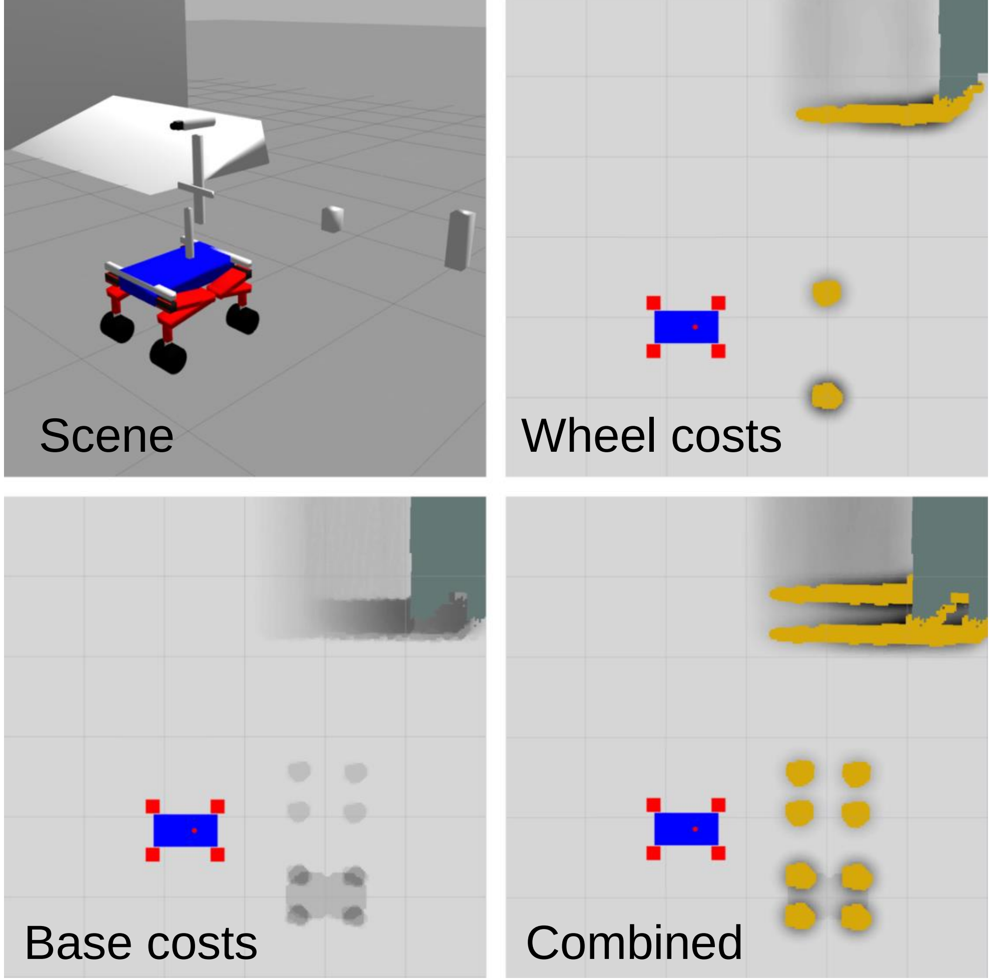
3D Map



Considering Robot Footprint

- Costs for individual wheel pairs from height differences
- Base costs
- Non-linear combination yields 3D (x, y, θ) cost map

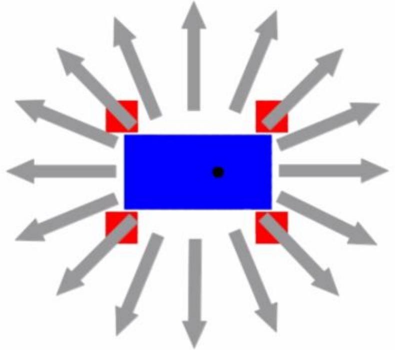
[Klamt and Behnke, IROS 2017]



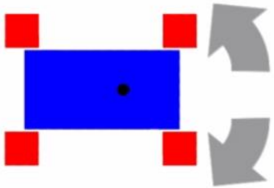
universitätbonn ais

3D Driving Planning (x, y, θ): A*

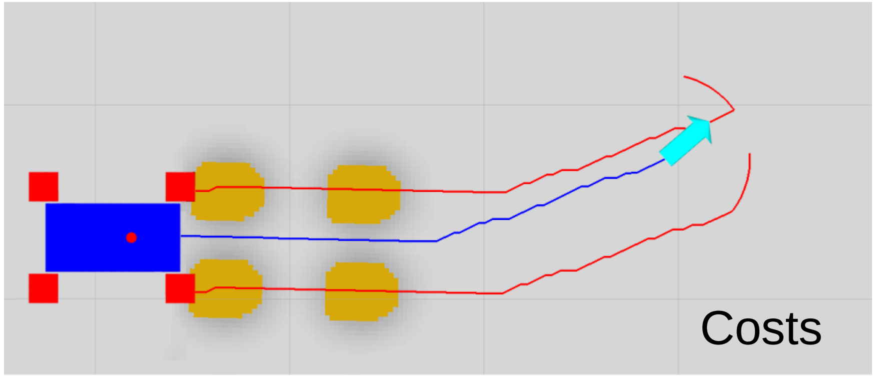
- 16 driving directions

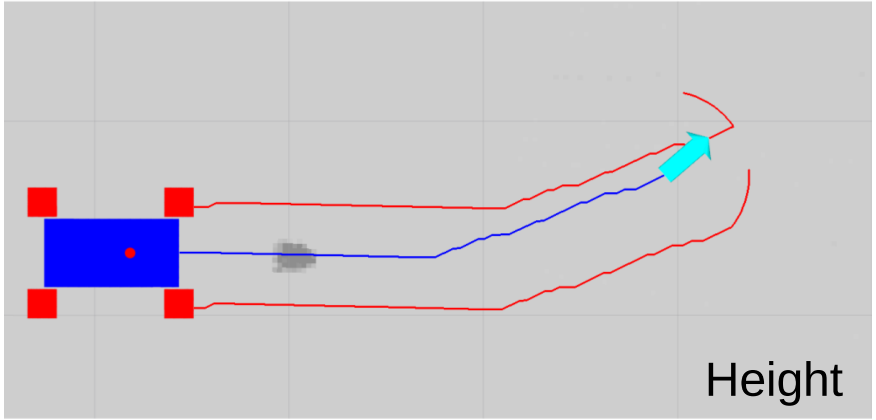


- Orientation changes



=> Obstacle between wheels



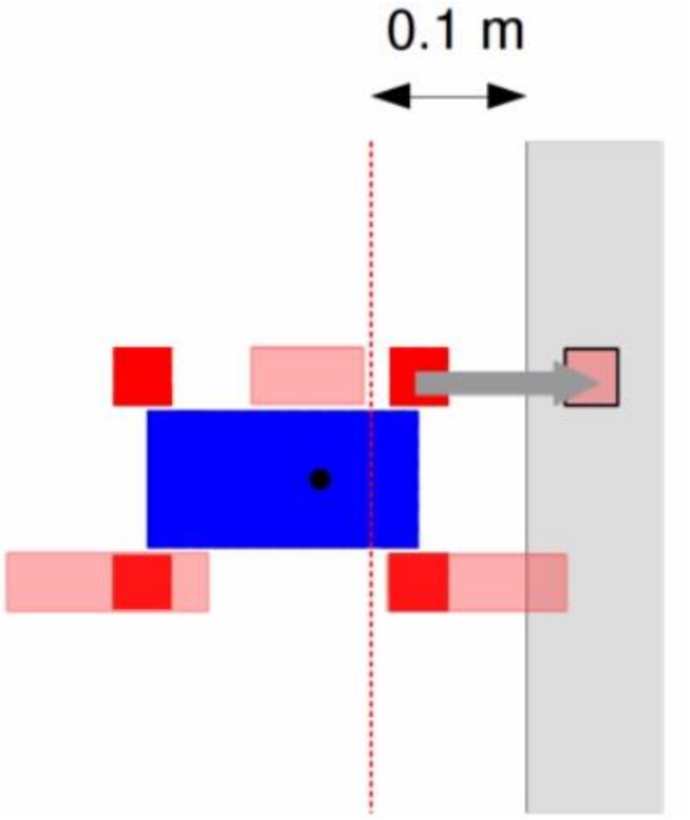


[Klamt and Behnke, IROS 2017]

universitätbonn ais

Making Steps

- If not drivable obstacle in front of a wheel
- Step landing must be drivable
- Support leg positions must be drivable

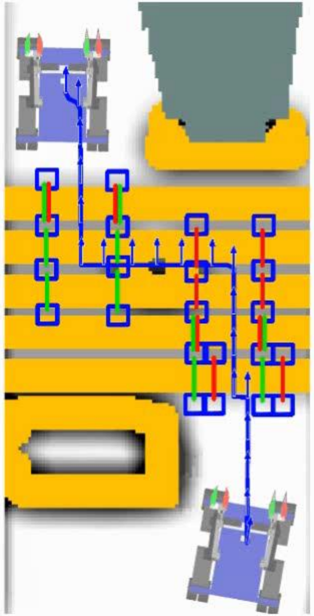


[Klamt and Behnke: IROS 2017]

Expanding Abstract Steps to Detailed Motion Sequences

[Klamt and Behnke: IROS 2017]

Planning for Challenging Scenarios



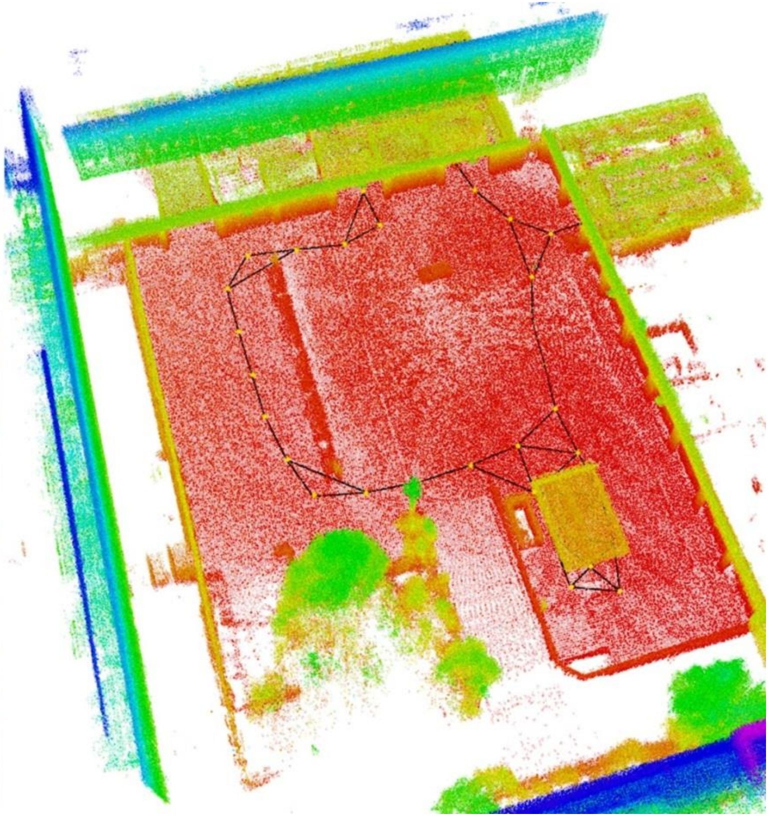
[Klamt and Behnke: IROS 2017]

New Sensor Head

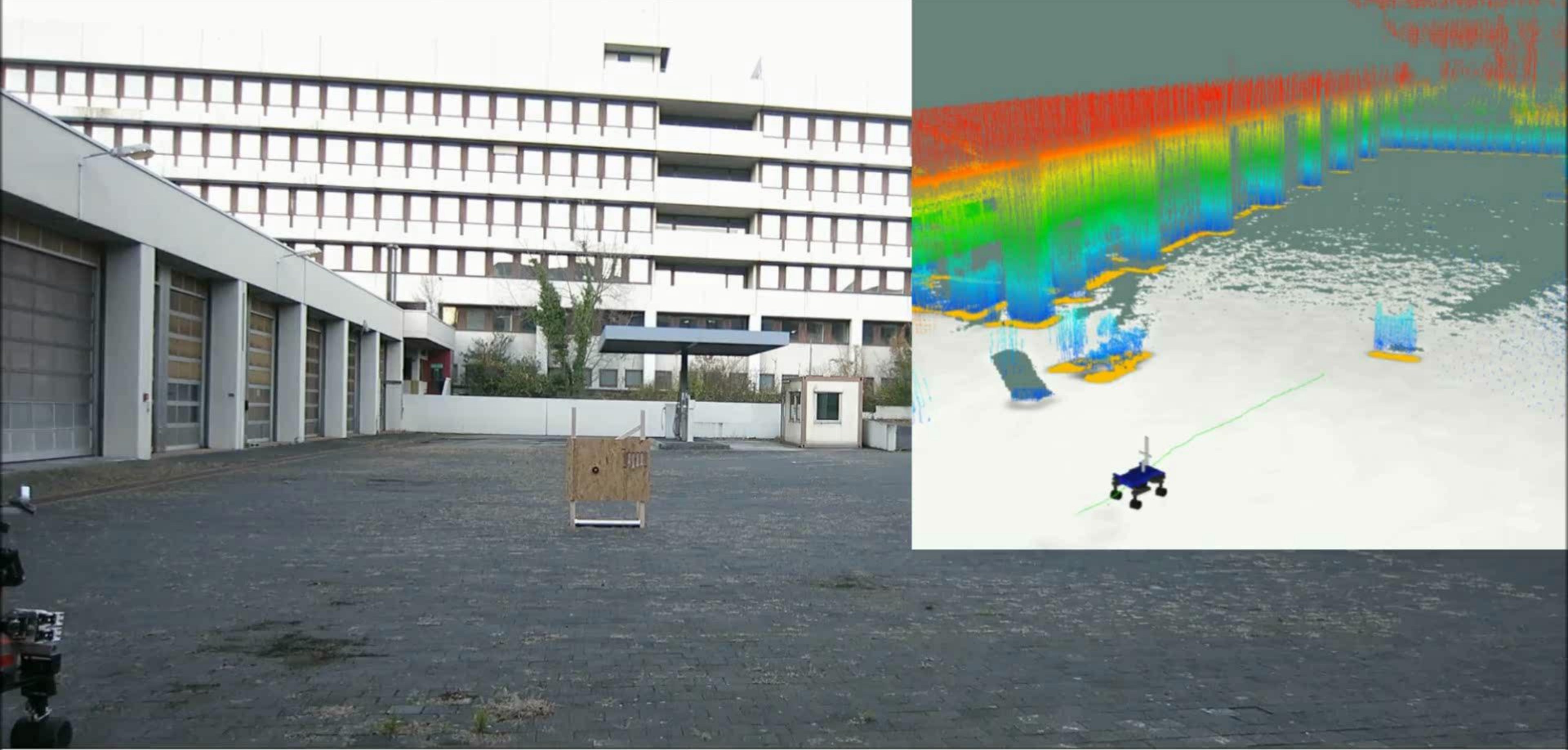
- Continuously rotating Velodyne Puck VLP-16
  - 300,000 3D points/s
  - 100 m range
  - Spherical field of view
- Three wide-angle color cameras (total FoV 210×103°)
- Kinect V2 RGB-D camera on pan-tilt unit

universitätbonn ais

3D Map of Indoor+Outdoor Scene



[Droeschel et al., Robotics and Autonomous Systems 2017]



Navigation in allocentric laser map (colored points)

Amazon Picking Challenge

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

[Schwarz et al. ICRA 2017]

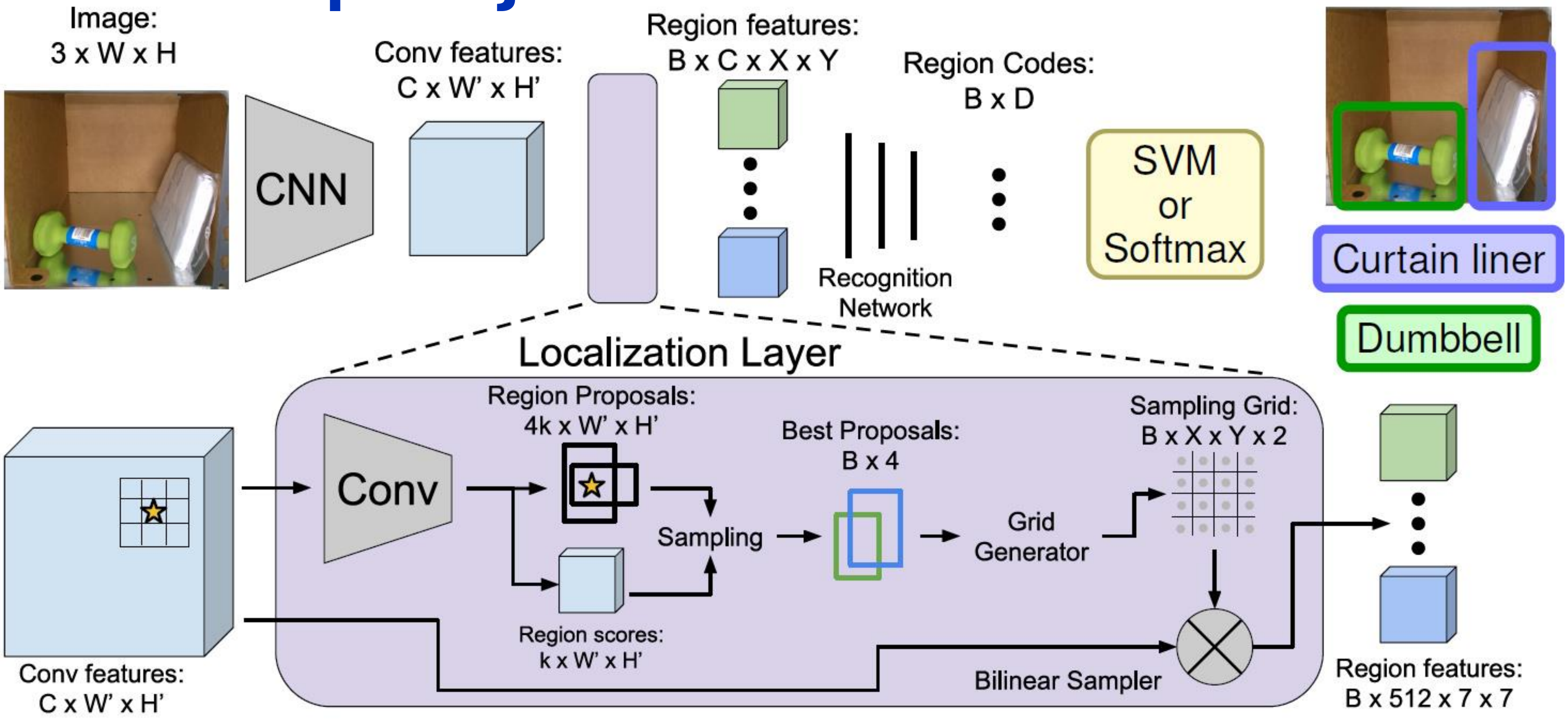
Deep Learning Semantic Segmentation

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

[Schwarz et al. ICRA 2017]

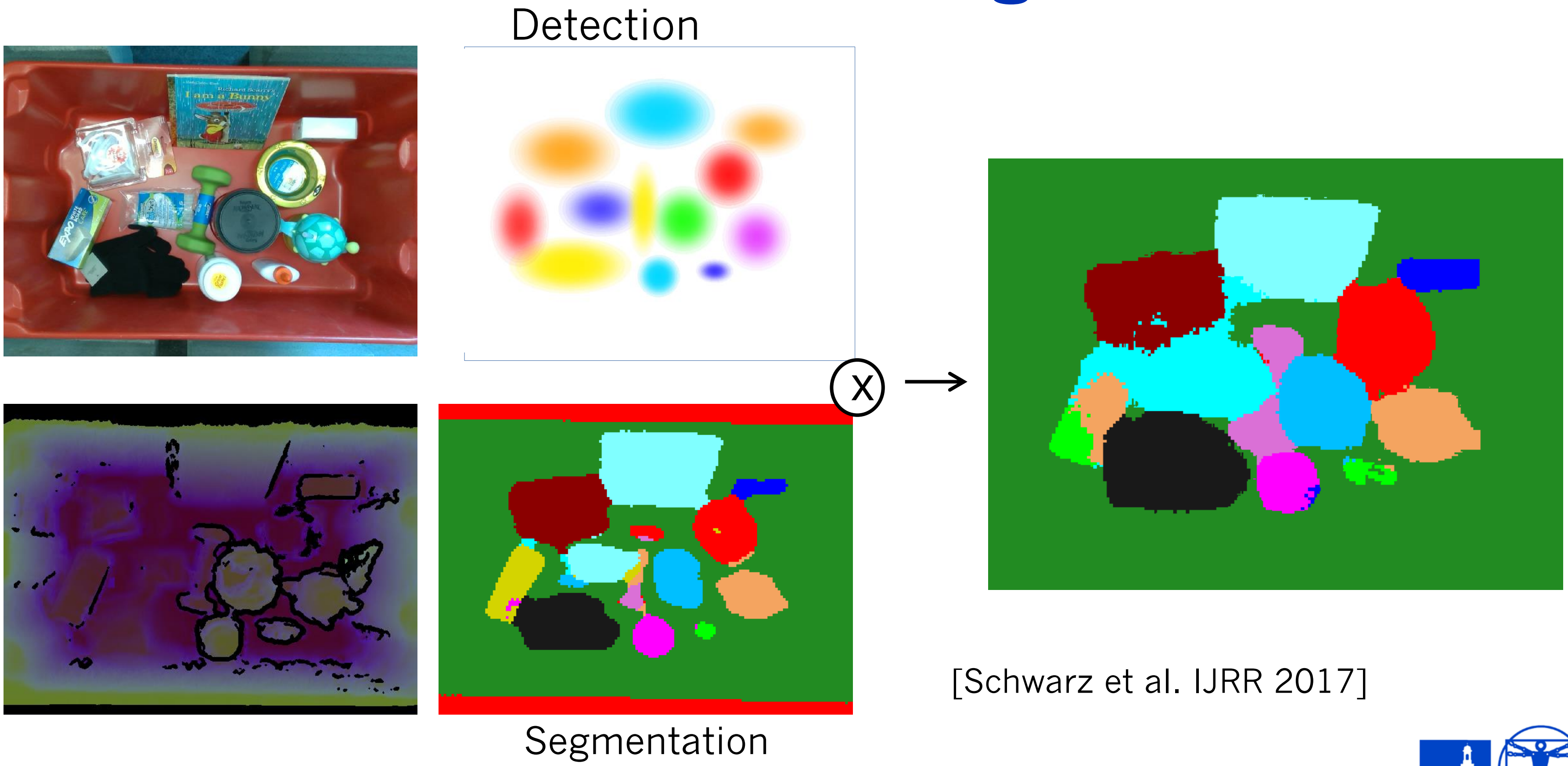
universitätbonn ais

DenseCap Object Detection



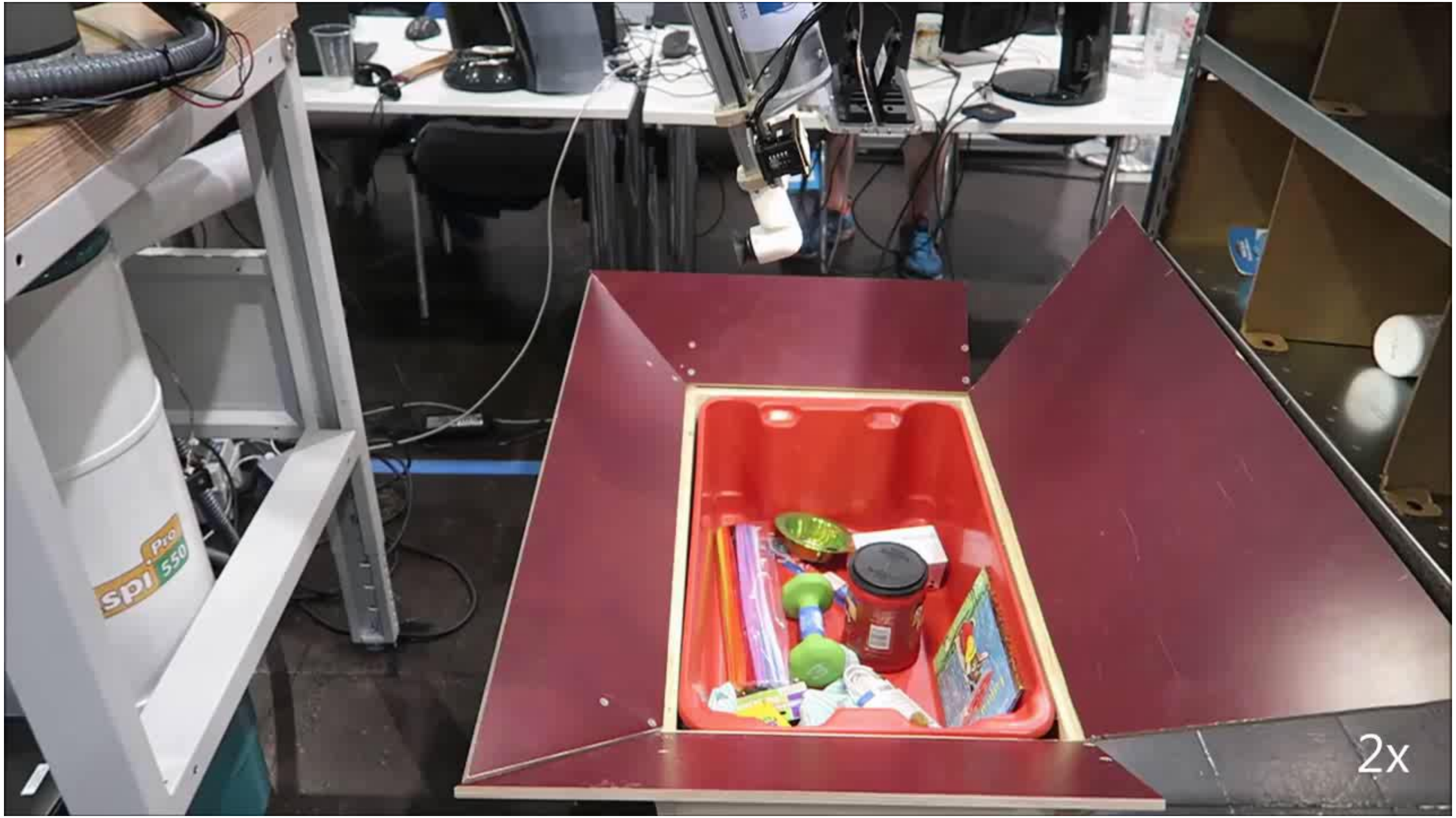
[Schwarz et al. ICRA 2017] [Johnson et a CVPR 2016]

Combined Detection and Segmentation



[Schwarz et al. IJRR 2017]

Stowing



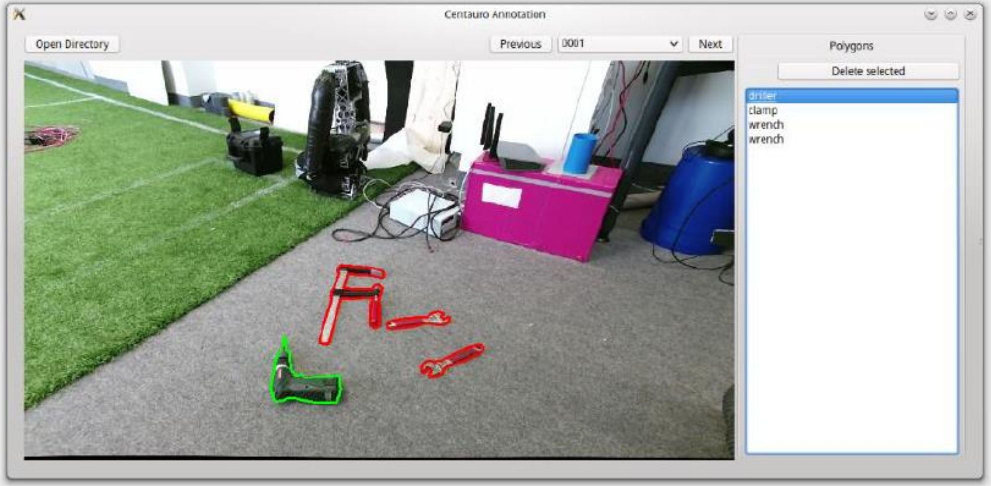
Picking

NimbRo Picking APC 2016 Results

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

[Schwarz et al. IJRR 2017]

CENTAURO Workspace Perception Data Set

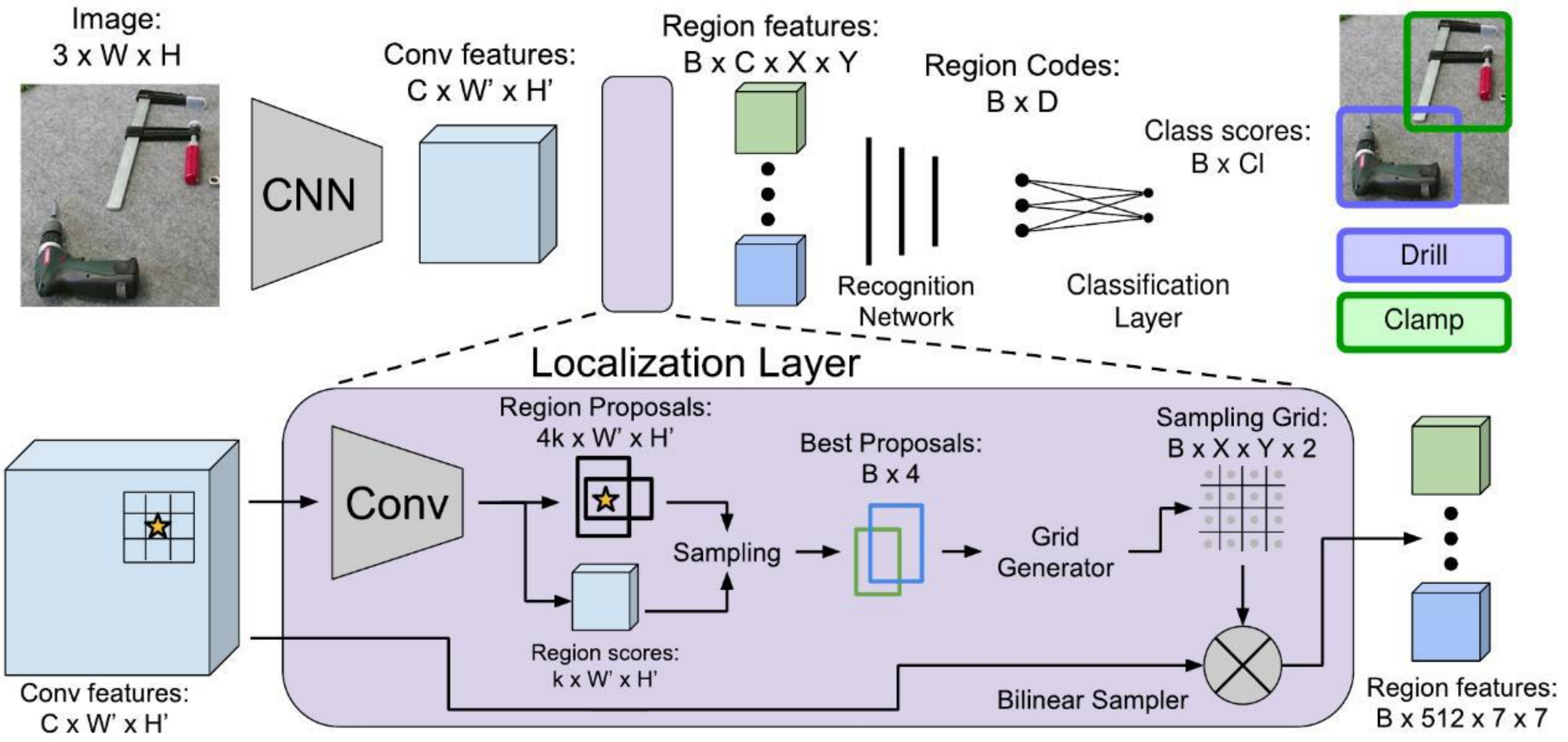


129 frames, 6 object classes

https://www.centauro-project.eu/data_multimedia/tools_data

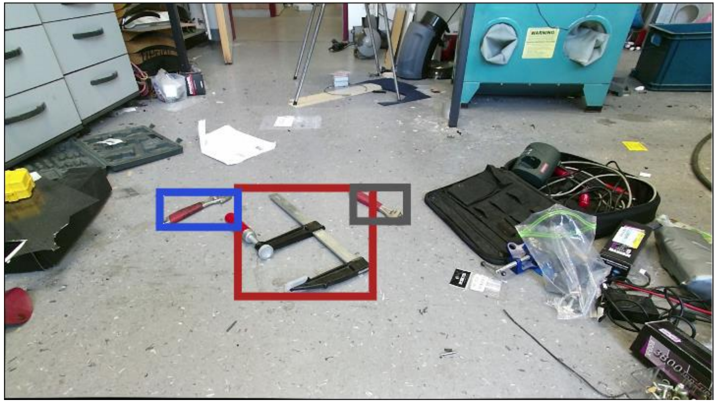
universitätbonn ais

Deep Learning Object Detection



[Johnson et al. 2015]

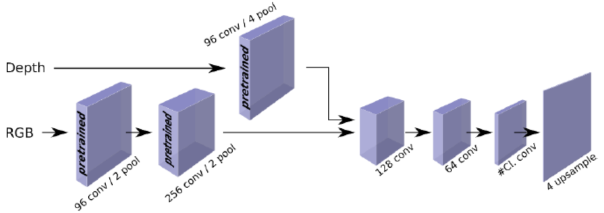
Detection of Tools



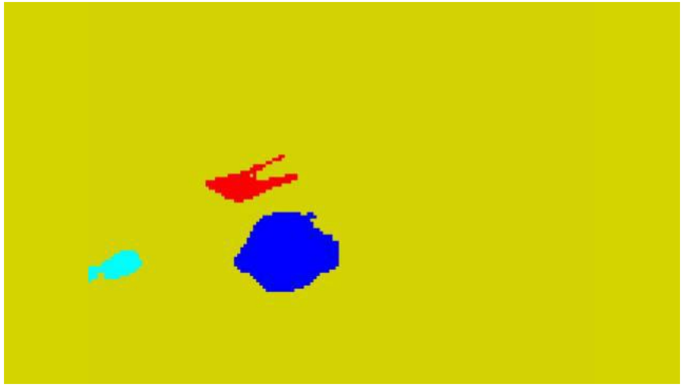
[Schwarz et al. IJRR 2017]

Semantic Segmentation

- Deep CNN



[Husain et al. RA-L 2016]



Pixel-wise accuracy:

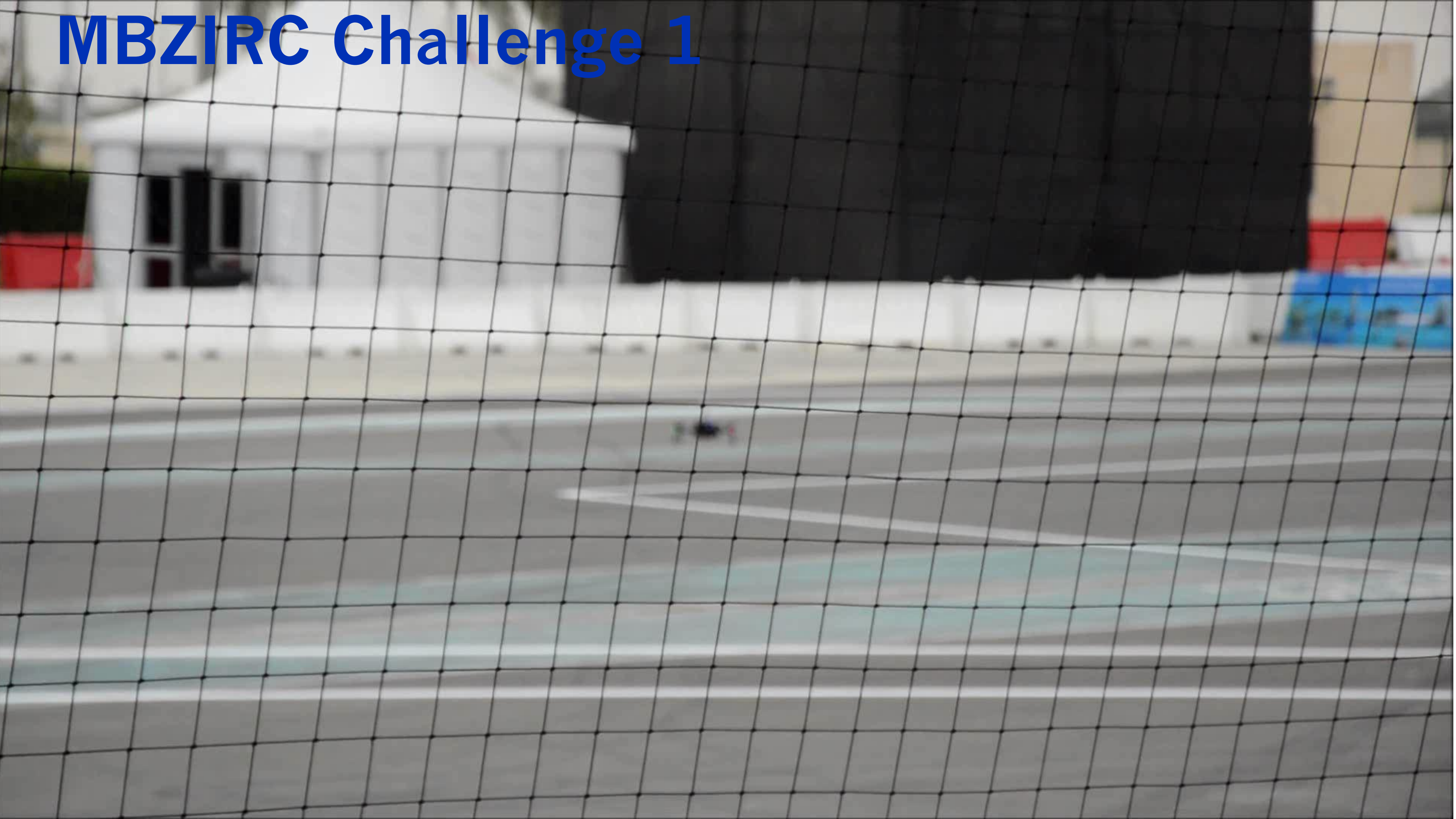
Clamp	Door handle	Driller	Extension	Stapler	Wrench	Background	Mean
0.727	0.751	0.769	0.889	0.775	0.734	0.992	0.805

MBZIRC Challenge 2

universitätbonn ais

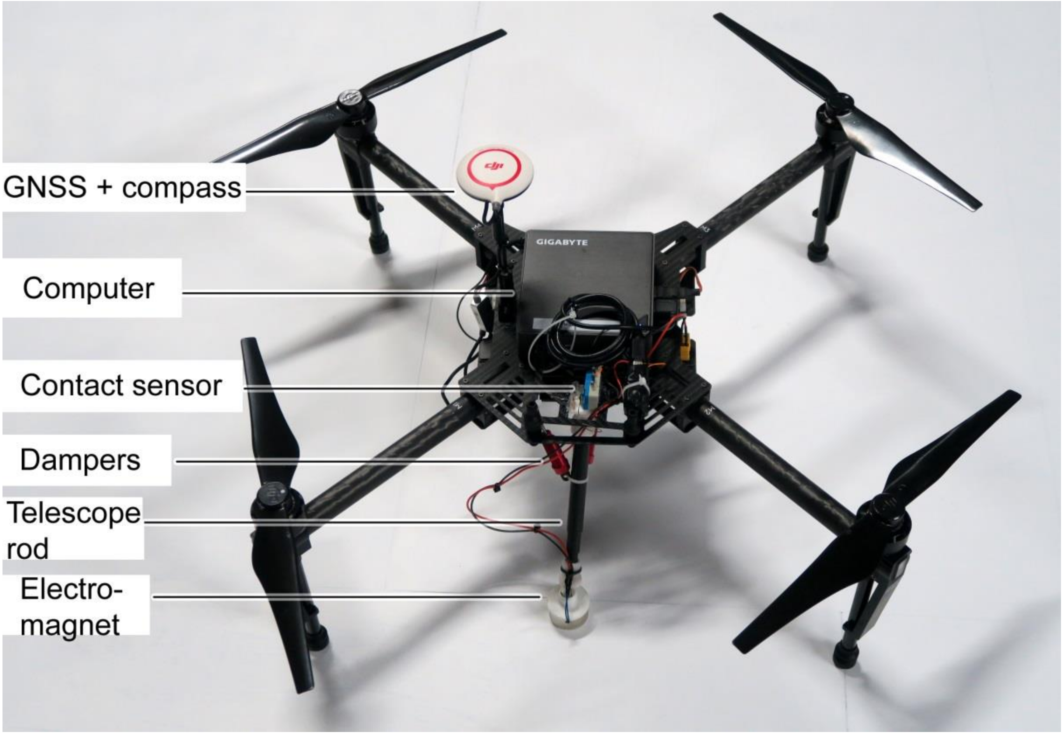
Wrench Selection: Detection of Tool Ends

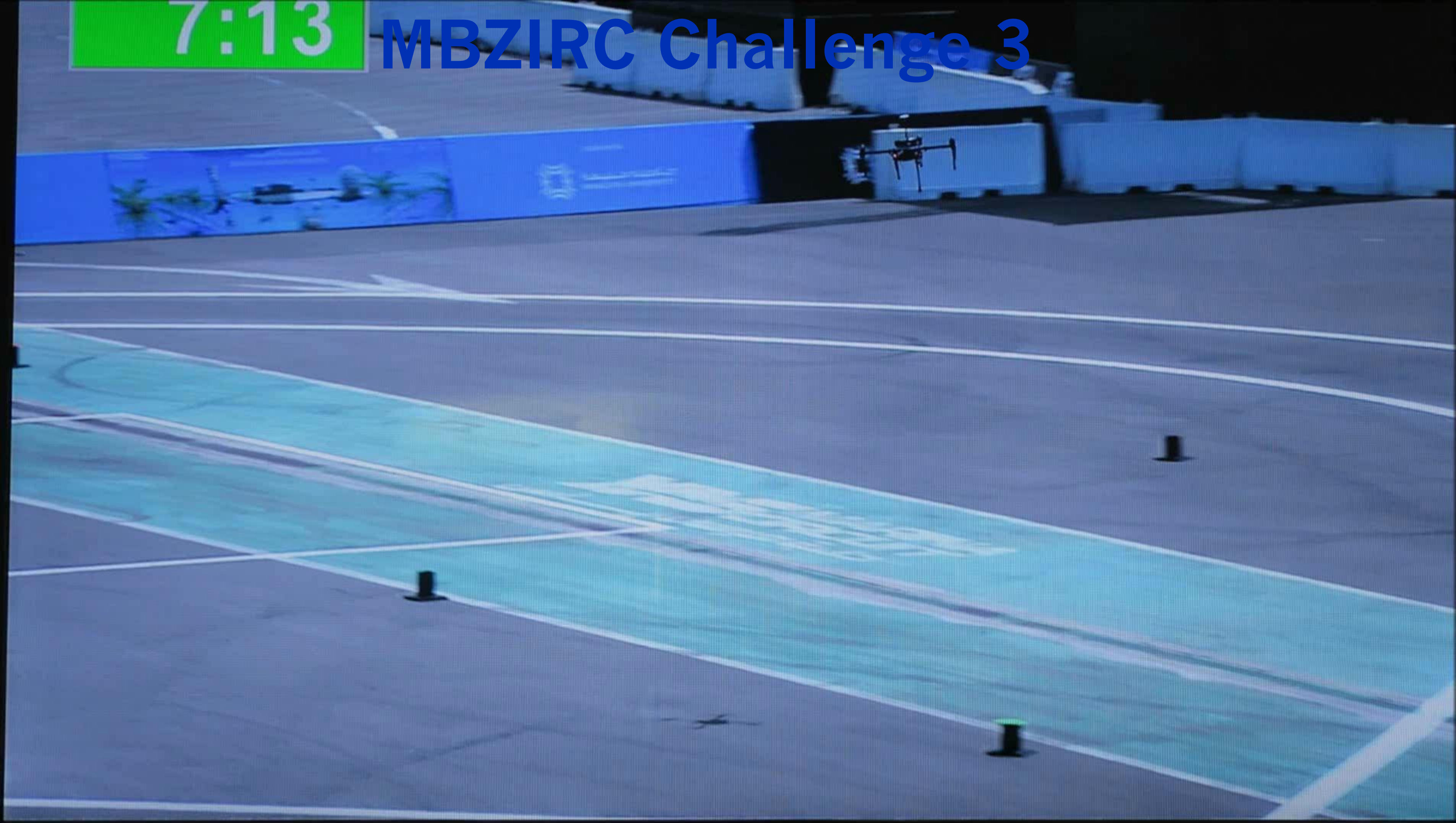
MBZIRC Challenge 1



Picking Copter DJI Matrice 100

- Wide-angle downward looking color camera
- Electromagnetic gripper
- Laser-distance sensor to ground
- Dual-core PC



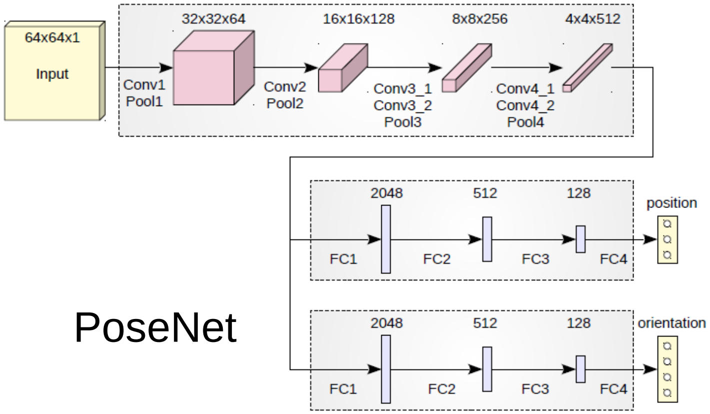

7:13
MBZIRC Challenge 3

MBZIRC Team NimbRo

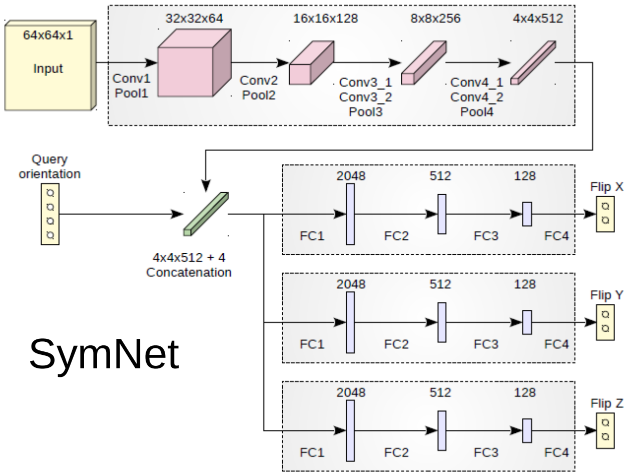
EuRoC C1 Robolink Feeder: Bin Picking

Part Pose Estimation

- Two convolutional neural networks

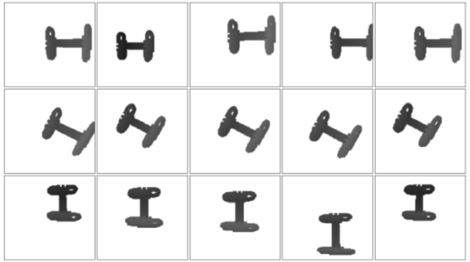


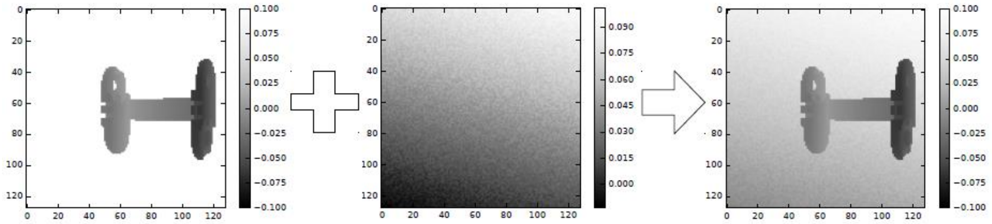
PoseNet



SymNet

- Training with synthetic depth images





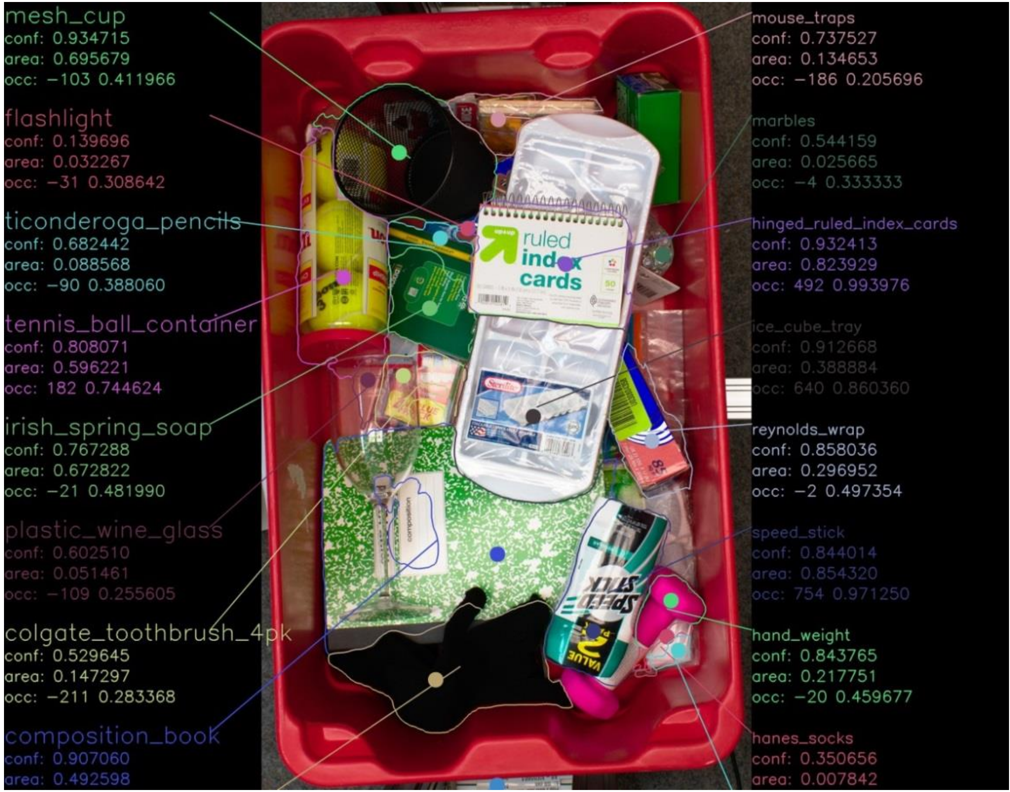
[Koo et al. CASE 2017]

universitätbonn ais

Robolink Feeder: Regrasping and Placing

Amazon Robotics Challenge 2017

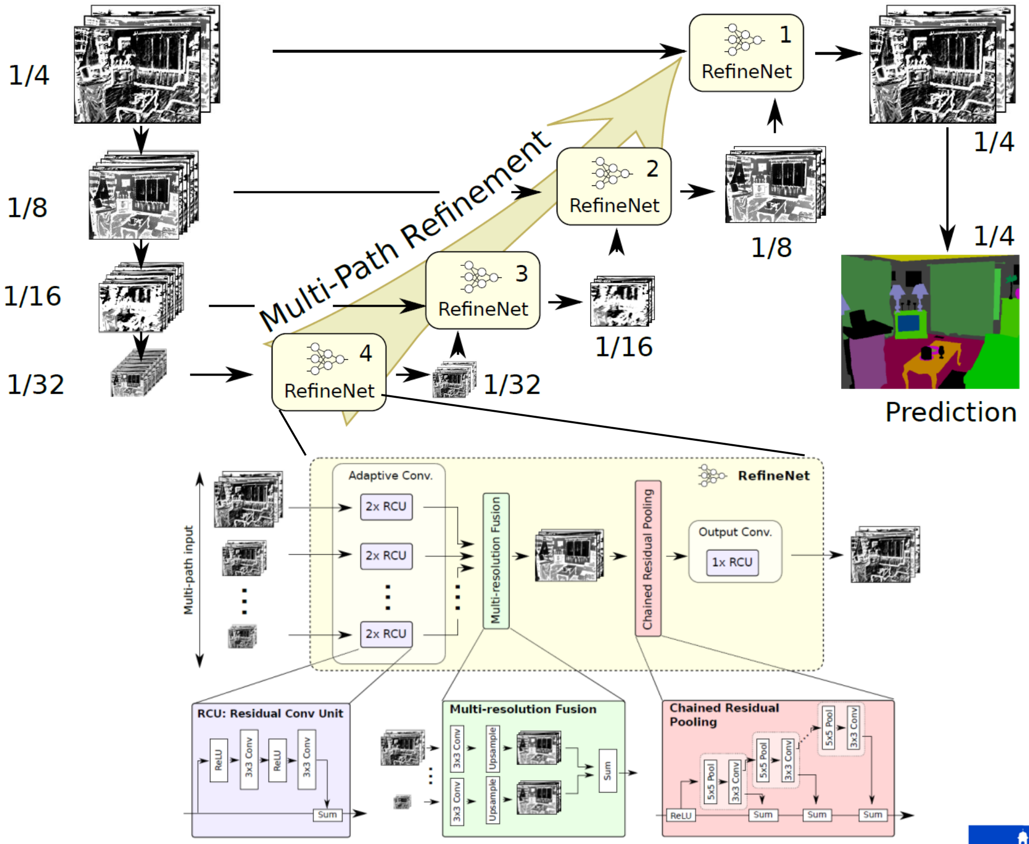
- Quick learning of novel objects
- Training with rendered scenes



RefineNet

[Lin et al. CVPR 2017]

- Increase reso-lution by using features from the higher resolution
- Corse-to-fine semantic segmentation

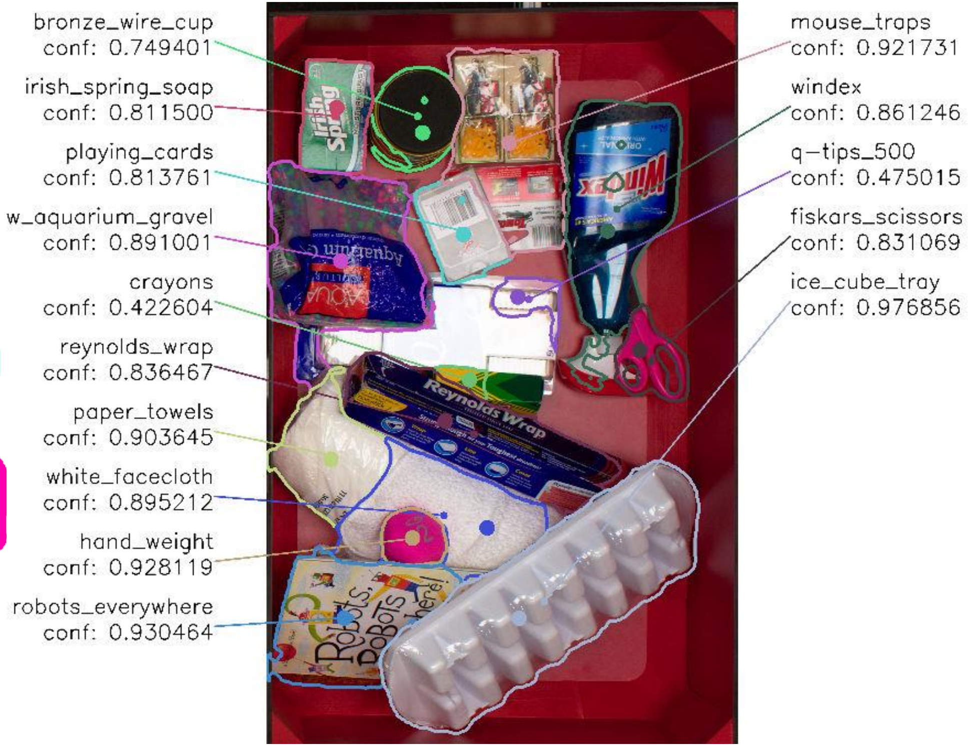


Object Capture and Scene Rendering

- Turn table + DLSR

Rendered scenes

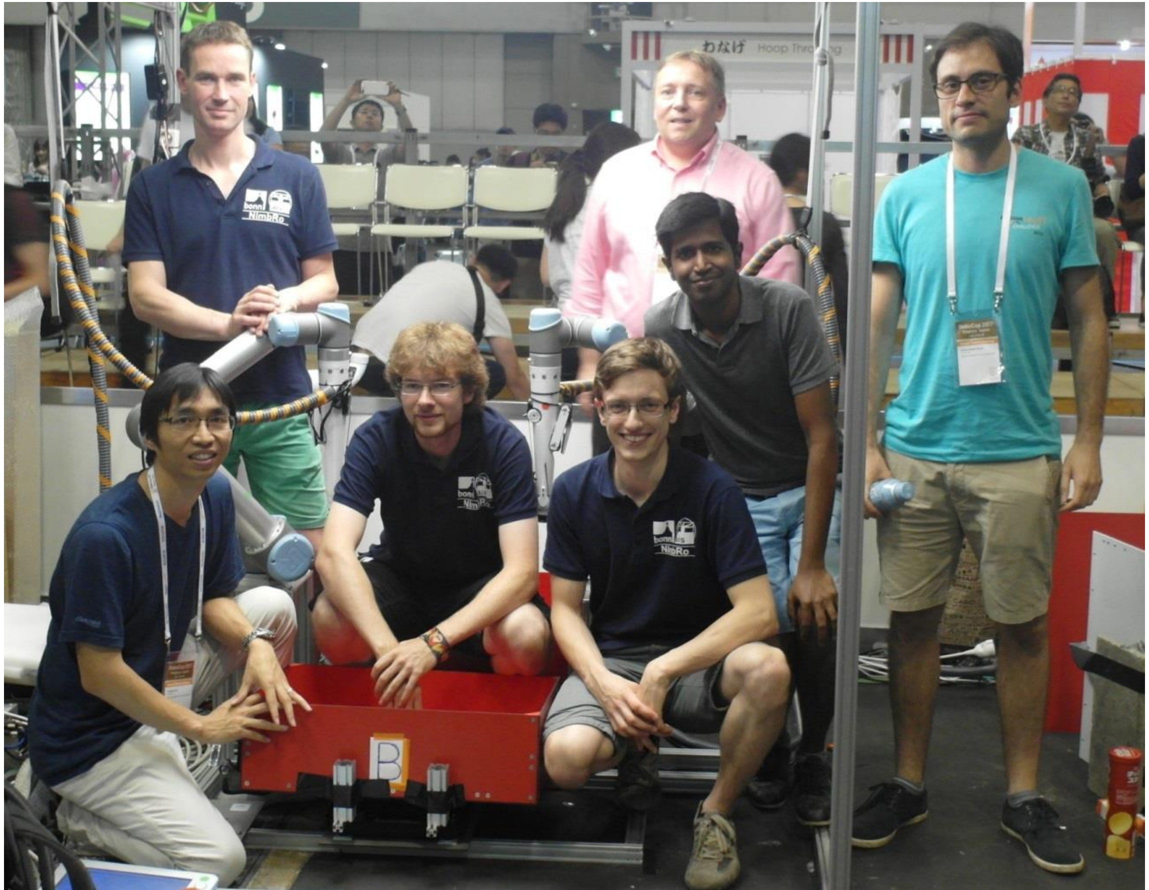
ARC 2017 Perception Example



Amazon Robotics Challenge 2017 Final

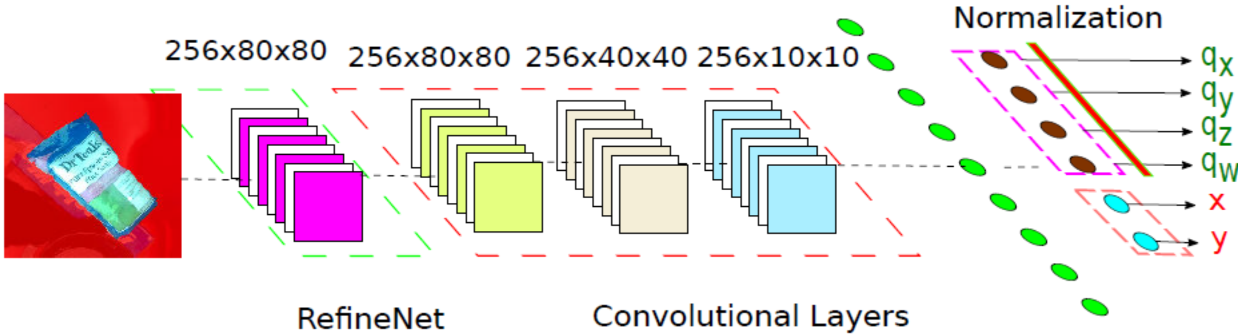
NimbRo Picking 2017 Team

- 2nd place Pick
- 2nd place Stow-and-Pick Final



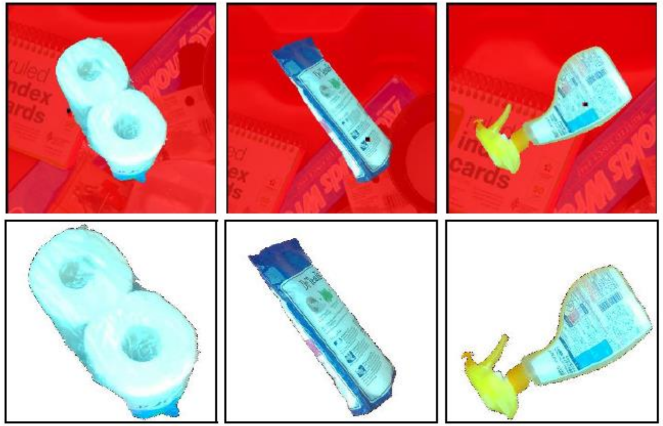
Object Pose Estimation

- Use upper layer of RefineNet as input
- Predict pose coordinates for one segment



Input

Predicted pose



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
- Synthetic training
- Many open problems, e.g. total scene understanding, incorporating physics, ...

Questions?