Generalizing Scene Parsing for Cluttered Bin Picking

Sven Behnke and Max Schwarz

University of Bonn
Computer Science Institute VI
Autonomous Intelligent Systems

Cognitive Robots @ AIS Bonn

- Focus on **cognitive** robot systems
- Equipped with numerous sensors and actuators
- Demonstration in complex scenarios

Soccer

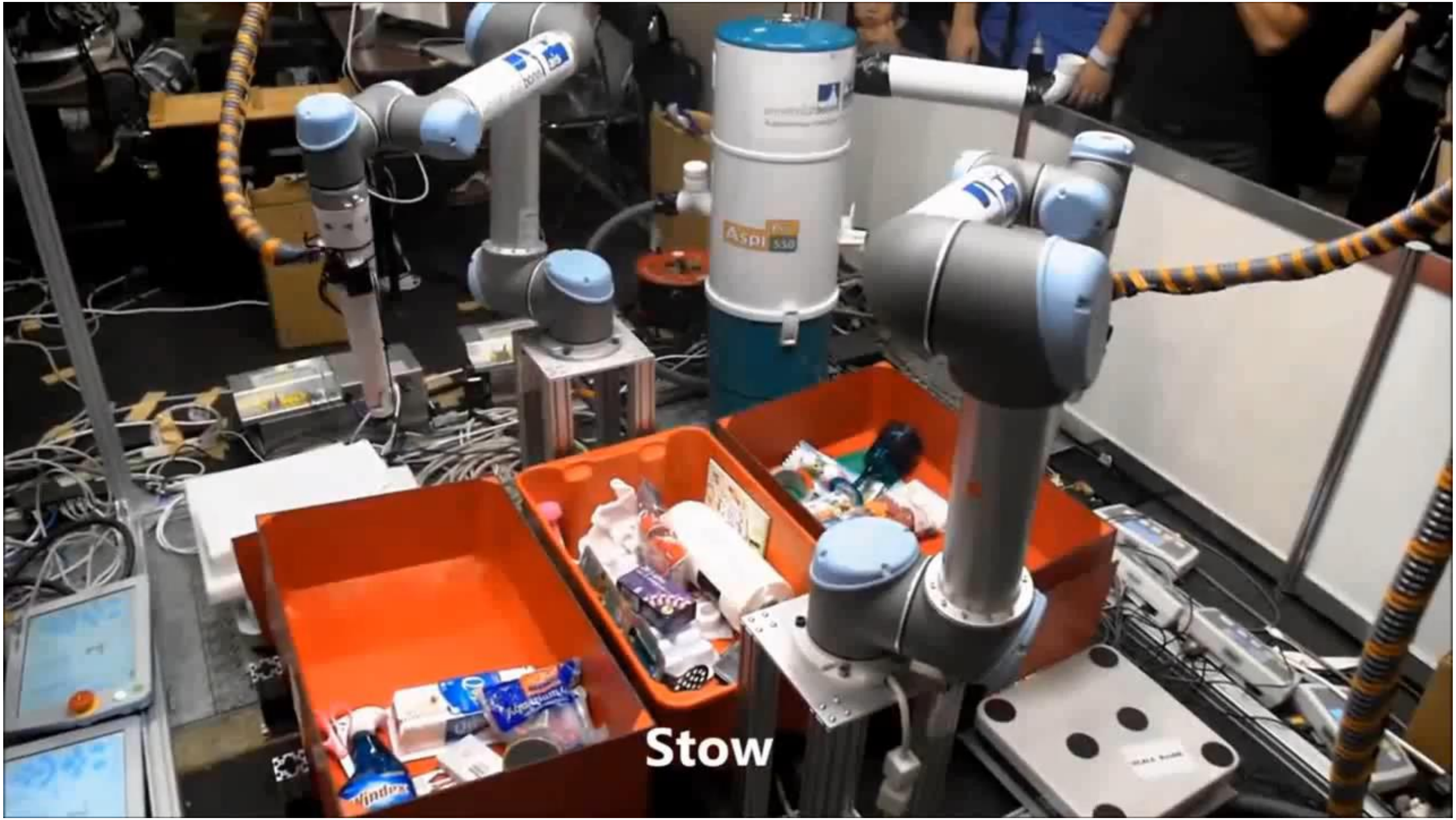
Domestic service

Mobile manipulation

Bin picking

Aerial inspection

NimbRo Picking @ Amazon Robotics Challenge 2017



[Schwarz et al. ICRA 2018]

Object Capture and Scene Rendering

- Turntable + DLSR camera

- Rendered scenes

[Schwarz et al. ICRA 2018]

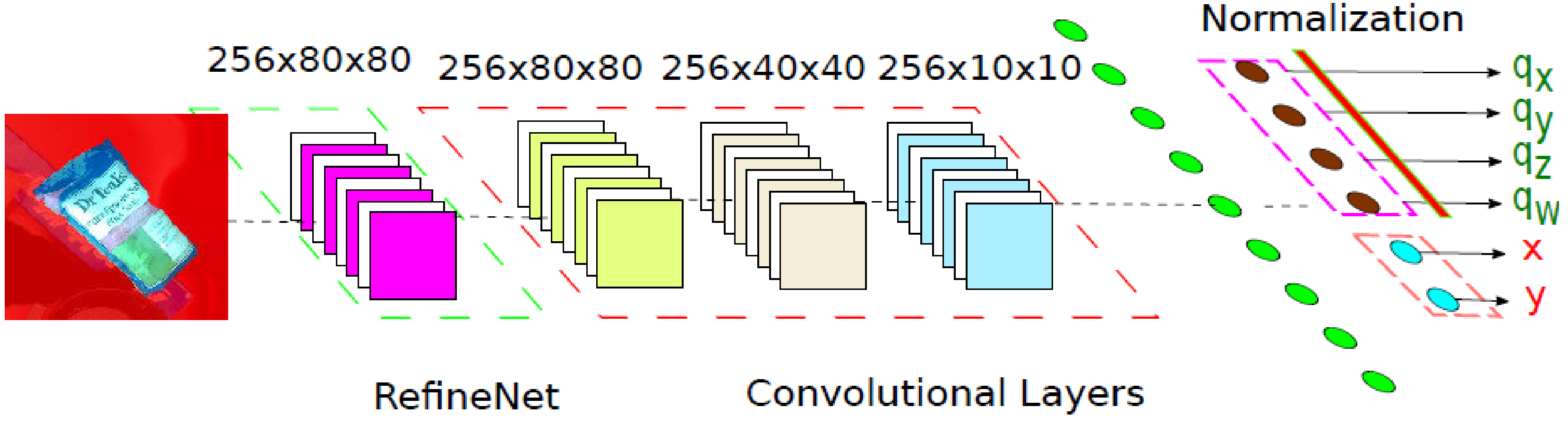
Semantic Segmentation

- Based on RefineNet [Lin et al. CVPR 2016]
- Suction grasp points in center of segments

[Schwarz et al. ICRA 2018]

Object Pose Estimation

- Cut out individual segments
- Use upper layer of RefineNet as input
- Predict pose coordinates
- Object mesh registration



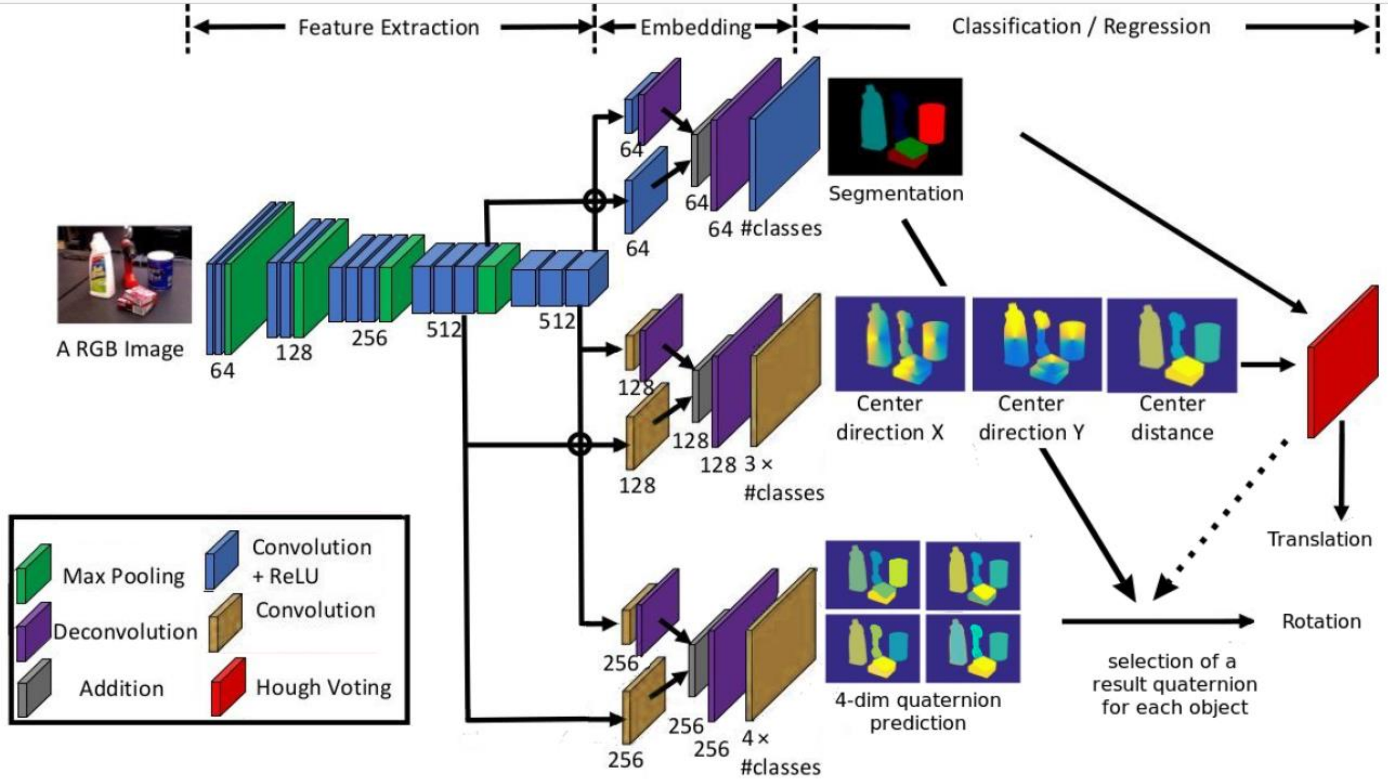
Input

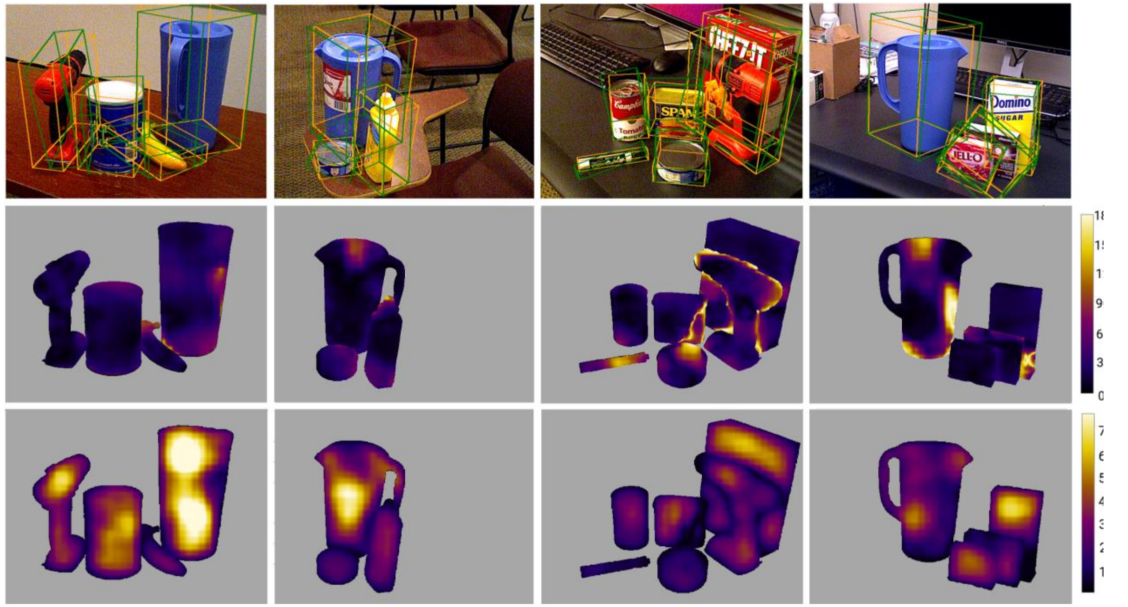
Predicted pose

[Schwarz et al. ICRA 2018, Periyasamy et al. IROS 2018]

Dense Convolutional 6D Object Pose Estimation

- Extension of PoseCNN [Xiang et al. RSS 2018]
- Dense prediction of object centers and orientations, without cutting out objects



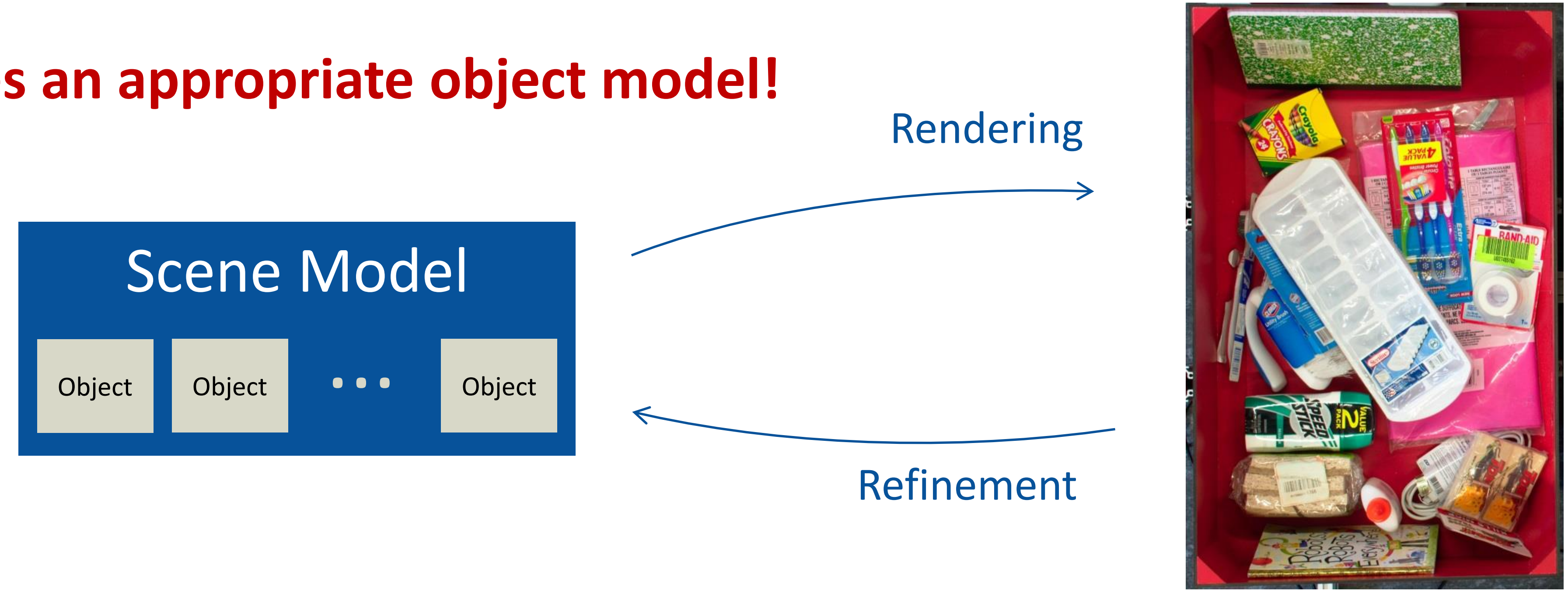


[Capellen 2019]

Towards Iterative Scene Parsing

- So far:
  - Instantaneous scene segmentation & rough pose estimation
- Goal:
  - Full scene **understanding**: Explain the measurements of the scene

Requires an appropriate object model!



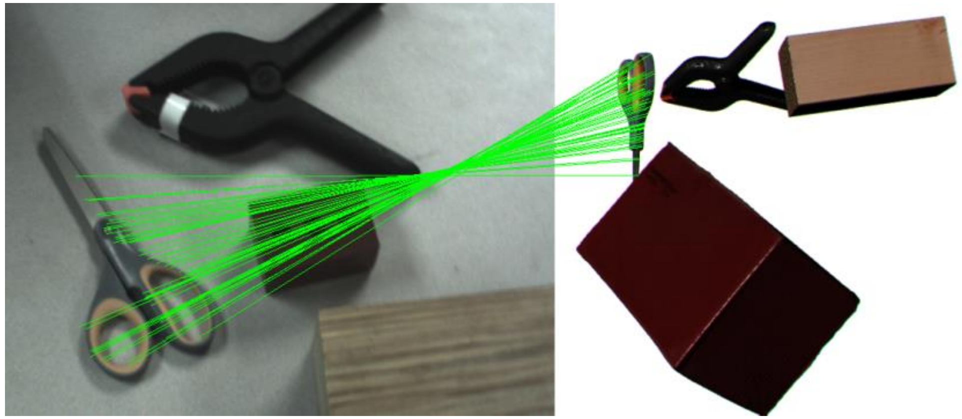
Training Data: From Turntable Captures to Textured Meshes

Fused & textured result

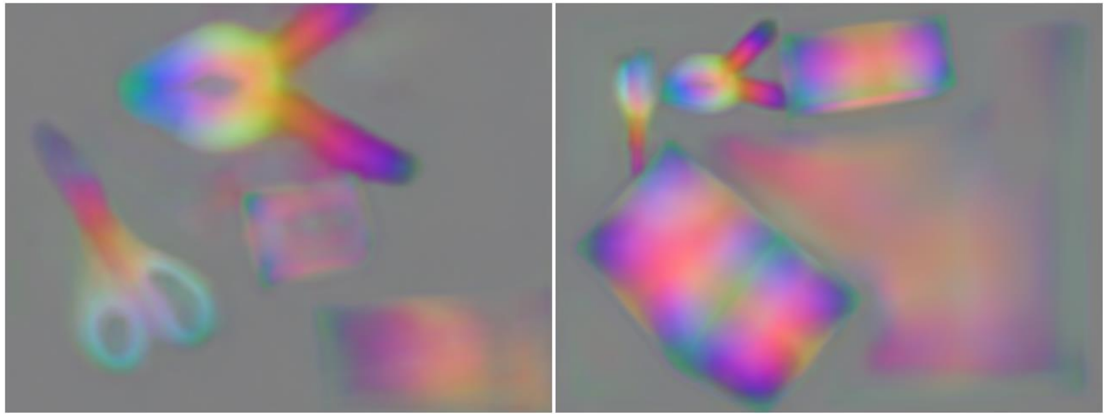
Improvement on method of Narayan et al. 2015, Publication pending

Self-Supervised Surface Descriptor Learning

- Feature descriptor should be constant under different transformations, viewing angles, and environmental effects such as lighting changes
- Descriptor should be unique to facilitate matching across different frames or representations
- Learn dense features using a contrastive loss [Schmidt et al. 2016]



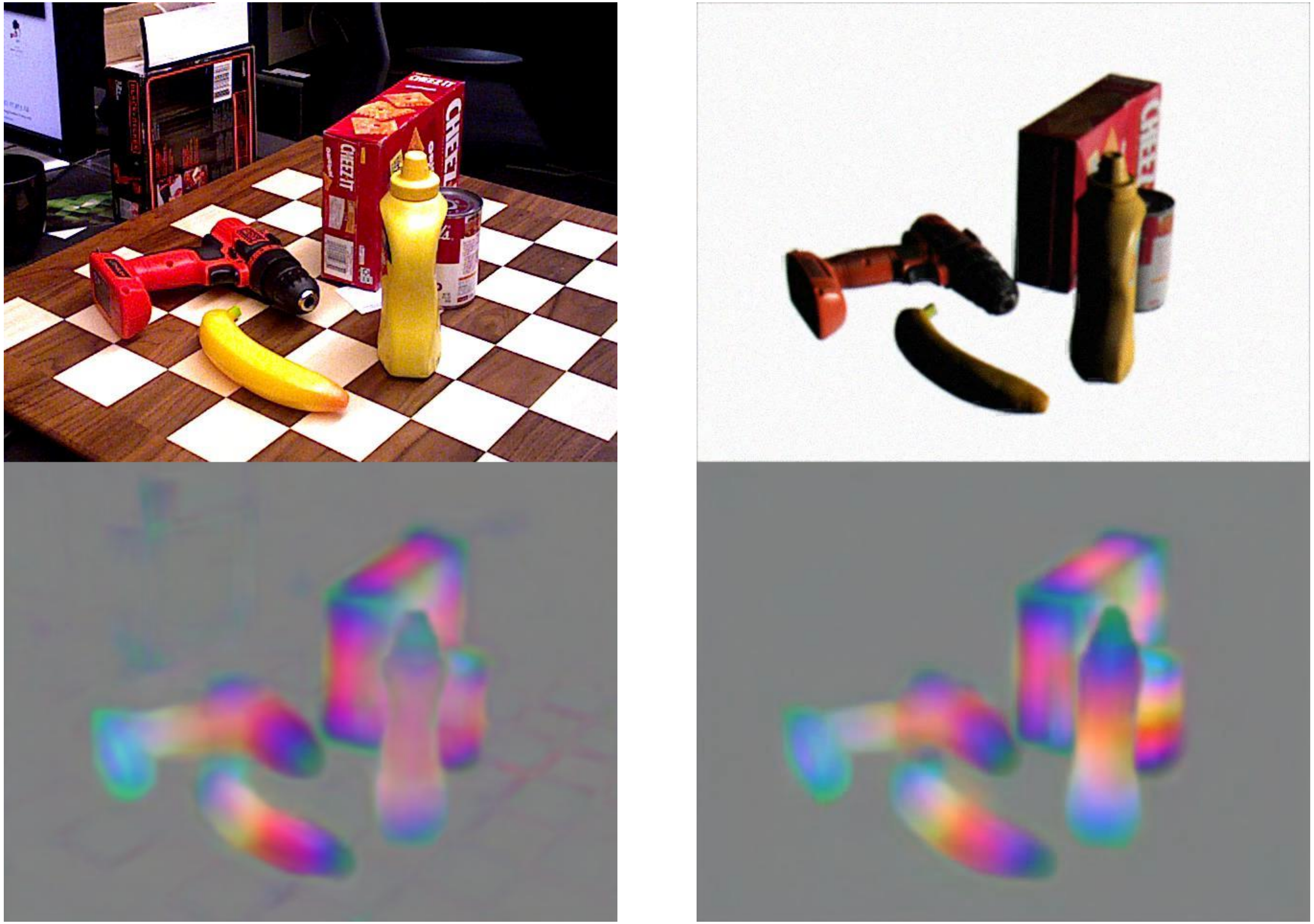
Known correspondences



Learned features

[Periyasamy, Schwarz, Behnke Humanoids 2019]

Learned Scene Abstraction



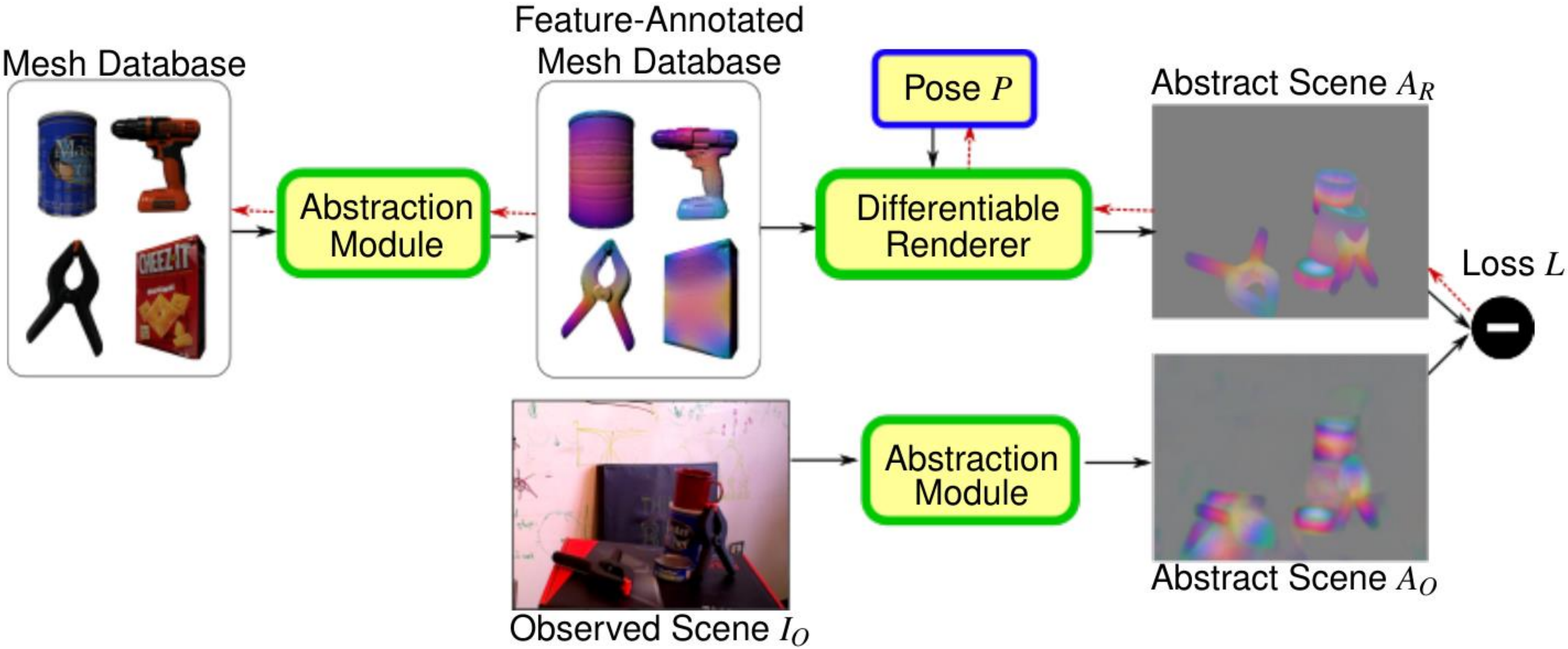
Descriptors as Texture on Object Surfaces

- Learned feature channels used as textures for 3D object models
- Used for 6D object pose estimation

[Periyasamy, Schwarz, Behnke Humanoids 2019]

Abstract Object Registration

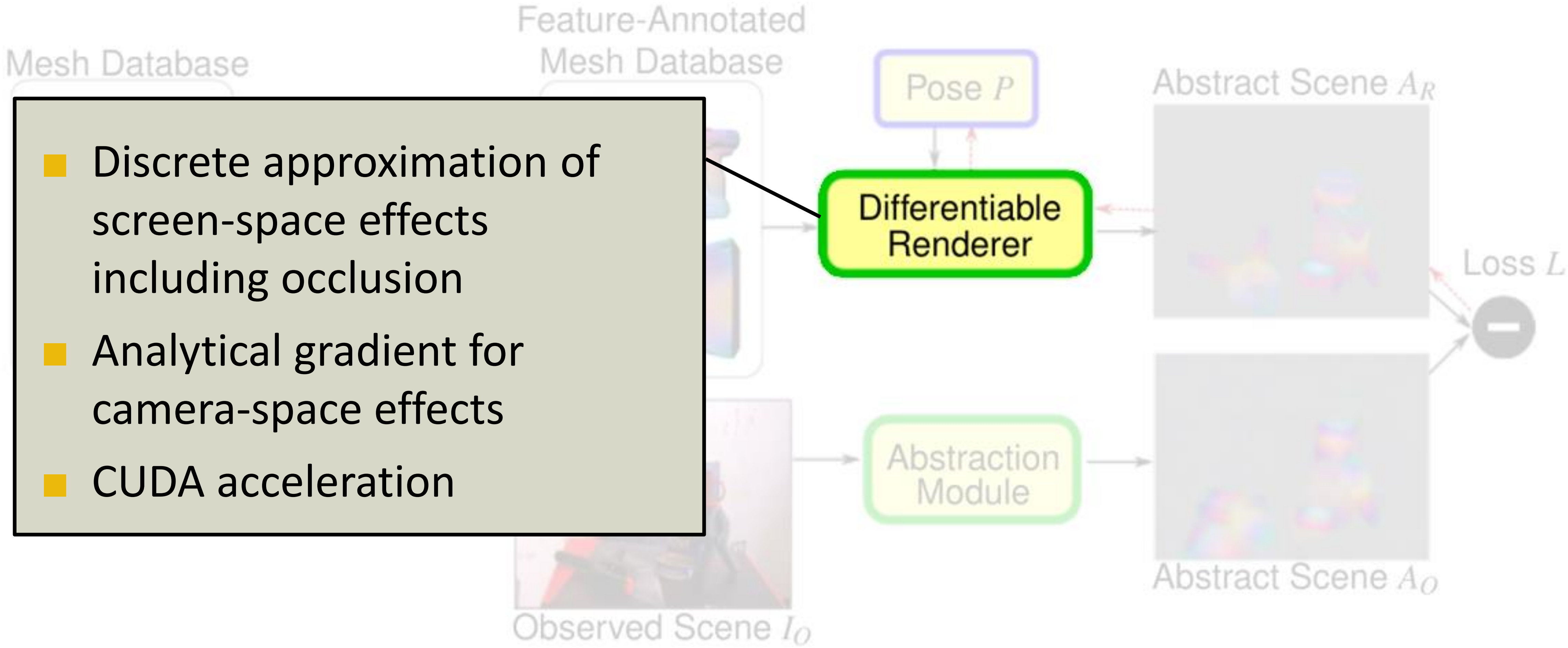
- Compare rendered and actual scene in feature space
- Adapt model pose by gradient descent



[Periyasamy, Schwarz, Behnke Humanoids 2019]

Abstract Object Registration

- Compare rendered and actual scene in feature space
- Adapt model pose by gradient descent



- Discrete approximation of screen-space effects including occlusion
- Analytical gradient for camera-space effects
- CUDA acceleration

[Periyasamy, Schwarz, Behnke Humanoids 2019]

Registration Examples

Evaluation on YCB-Video

- Consistent improvement on PoseCNN (Xiang et al. 2018) predictions
- Monocular (no depth)

	PoseCNN [5]		PoseCNN refined (ours)	
Object	ADD	ADD-S	ADD(Δ)	ADD-S(Δ)
master_chef_can	50.2	83.9	63.3(+13.1)	91.7(+7.8)
cracker_box	53.1	76.9	65.3(+12.2)	81.7(+4.9)
sugar_box	68.4	84.2	85.3(+16.9)	92.0(+7.8)
tomato_soup_can	66.2	81.0	59.4(−6.8)	79.9(−1.1)
mustard_bottle	81.0	90.4	86.5(+5.5)	92.3(+1.9)
tuna_fish_can	70.7	88.0	81.1(+10.4)	94.3(+6.3)
pudding_box	62.7	79.1	71.1(+8.4)	83.1(+4.1)
gelatin_box	75.2	87.2	81.5(+6.3)	89.1(+1.9)
potted_meat_can	59.5	78.5	63.7(+4.2)	80.3(+1.8)
banana	72.3	86.0	82.1(+9.8)	91.8(+5.8)
pitcher_base	53.3	77.0	85.1(+31.8)	92.7(+15.7)
bleach_cleanser	50.3	71.6	65.0(+14.7)	80.4(+8.9)
bowl	3.3	69.6	6.5(+3.1)	75.5(+5.9)
mug	58.5	78.2	65.9(+7.4)	84.0(+5.9)
power_drill	55.3	72.7	73.7(+18.4)	85.9(+13.2)
wood_block	26.6	64.3	45.5(+18.9)	73.3(+9.0)
scissors	35.8	56.9	40.0(+4.1)	58.6(+1.7)
large_marker	58.3	71.7	63.9(+5.6)	77.3(+5.6)
large_clamp	24.6	50.2	37.0(+12.4)	65.1(+15.0)
extra_large_clamp	16.1	44.1	25.4(+9.3)	63.7(+19.6)
foam_brick	40.2	88.0	43.3(+3.1)	90.8(+2.8)
ALL	53.7	75.8	62.8(+9.1)	82.4(+6.6)

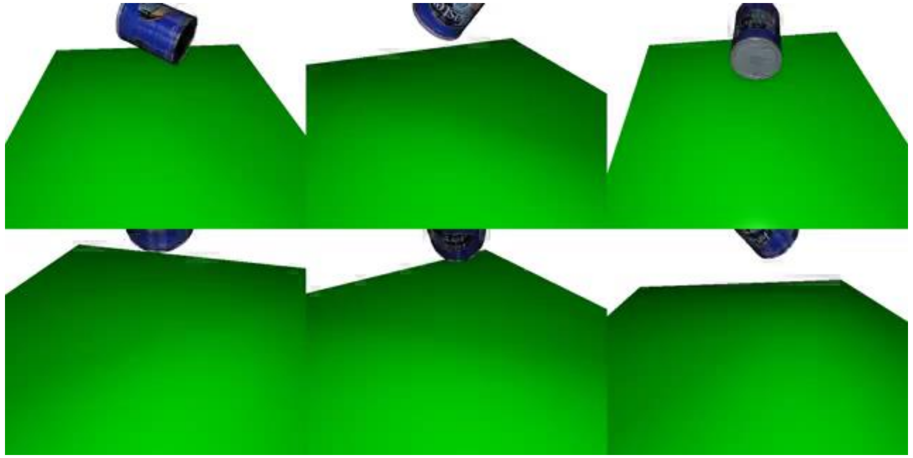

PoseCNN initialization

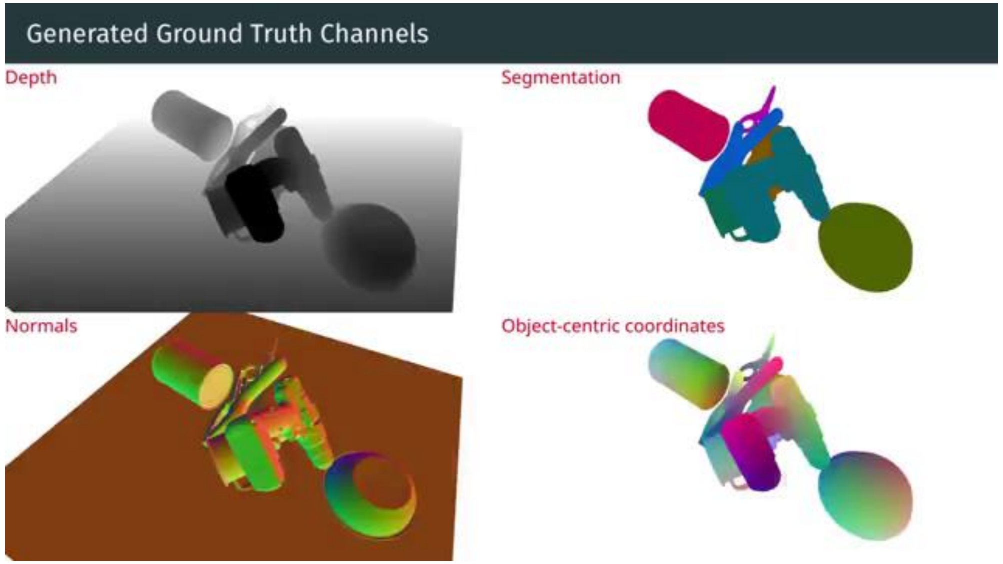

Optimization result

[Periyasamy, Schwarz, Behnke Humanoids 2019]

Learning from Synthetic Scenes

- Cluttered arrangements from 3D meshes
- Photorealistic scenes with randomized material and lighting including ground truth
- For online learning & render-and-compare
- Semantic segmentation on YCB Video Dataset
  - Close to real-data accuracy
  - Improves segmentation of real data

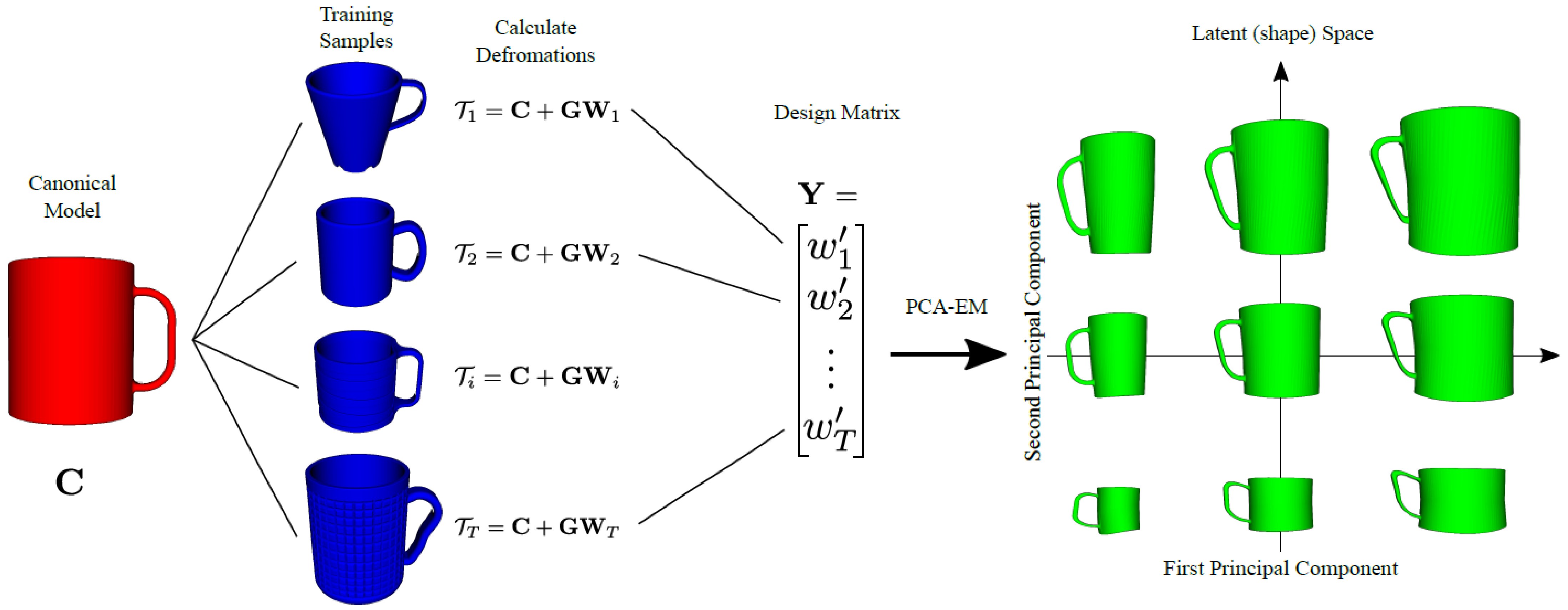




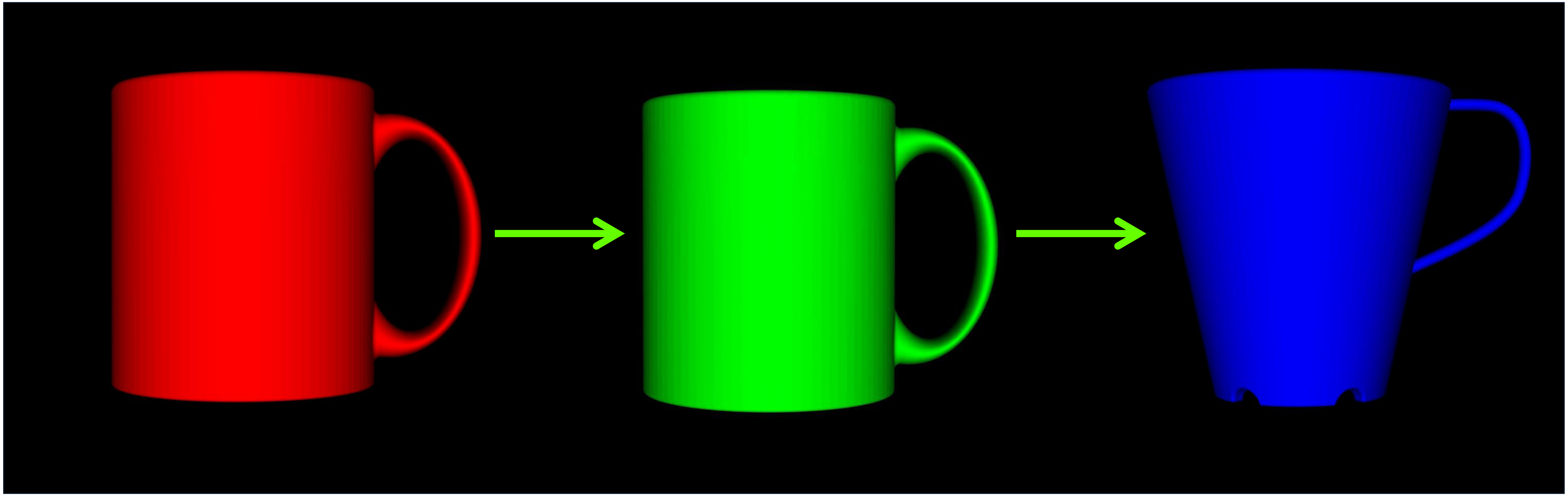
[Schwarz et al. 2020 (submitted)]

Learning a Latent Shape Space

- Non-rigid registration of instances and canonical model
- Principal component analysis of deformations



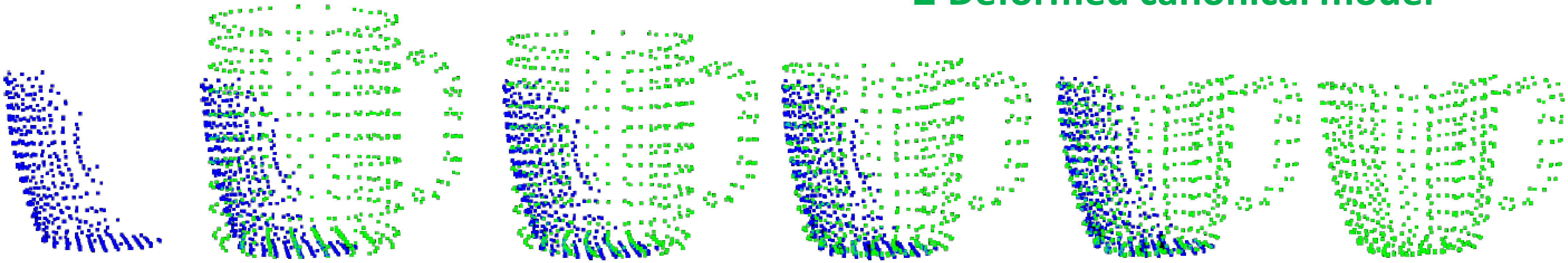
Interpolation in Shape Space

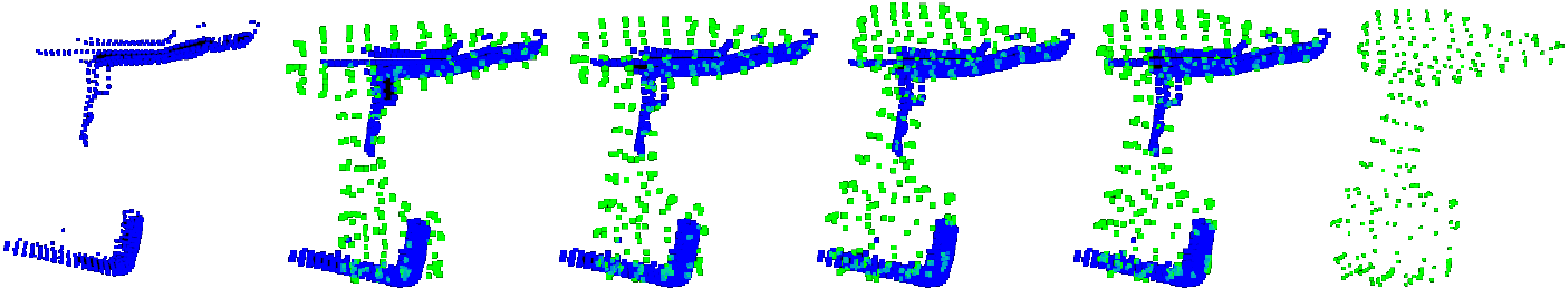


[Rodriguez and Behnke ICRA 2018]

Shape-aware Non-rigid Registration

- Partial view of novel instance
- Deformed canonical model



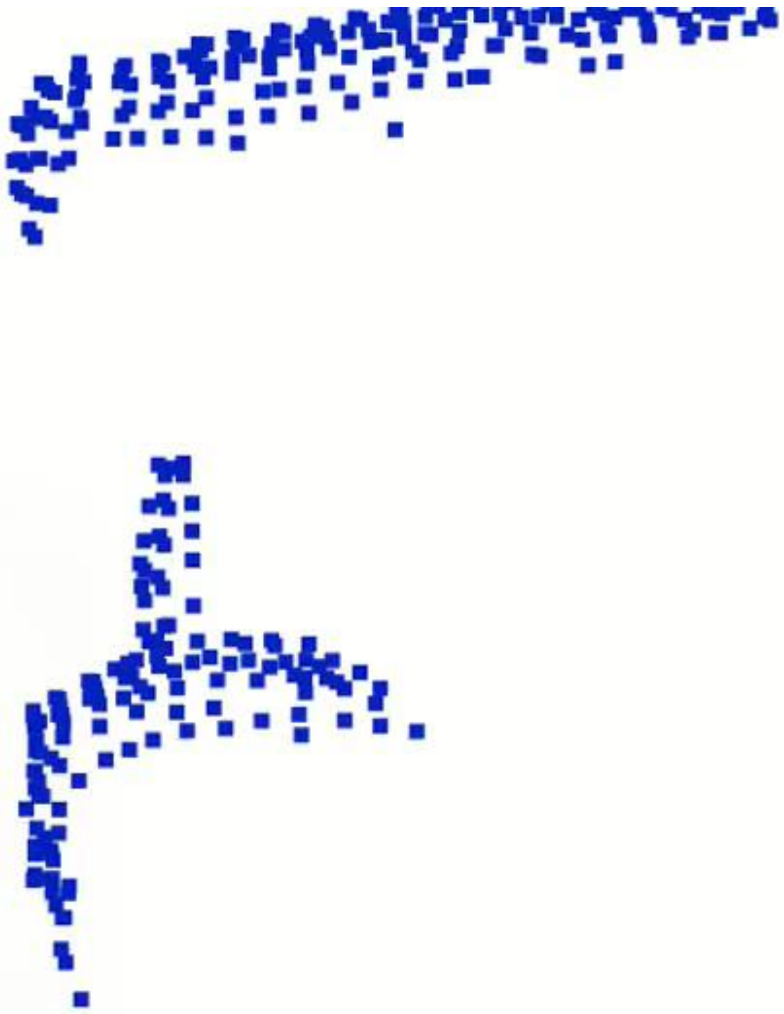


[Rodriguez and Behnke ICRA 2018]

Shape-aware Registration for Grasp Transfer

- Full point cloud

- Partial view

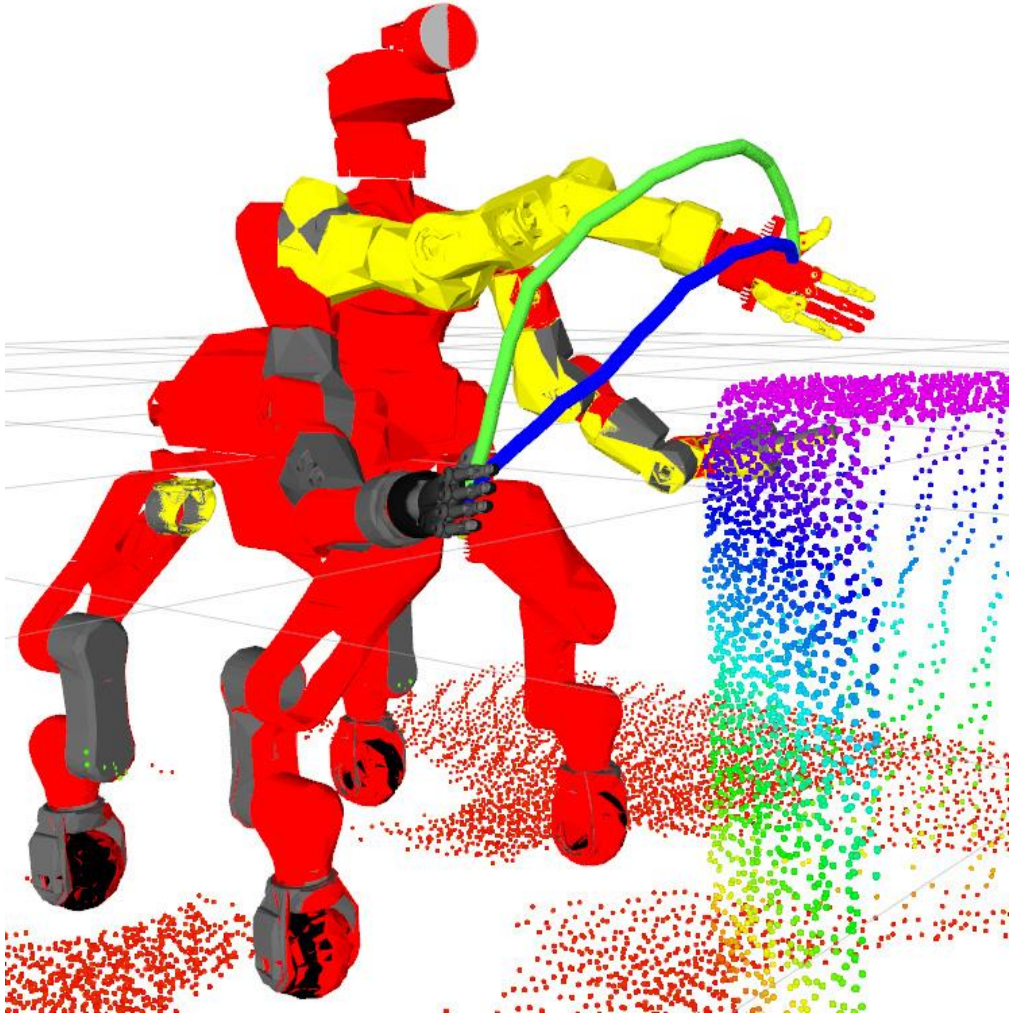


[Rodriguez and Behnke ICRA 2018]

Collision-aware Motion Generation

Constrained Trajectory Optimization:

- Collision avoidance
- Joint limits
- Time minimization
- Torque optimization



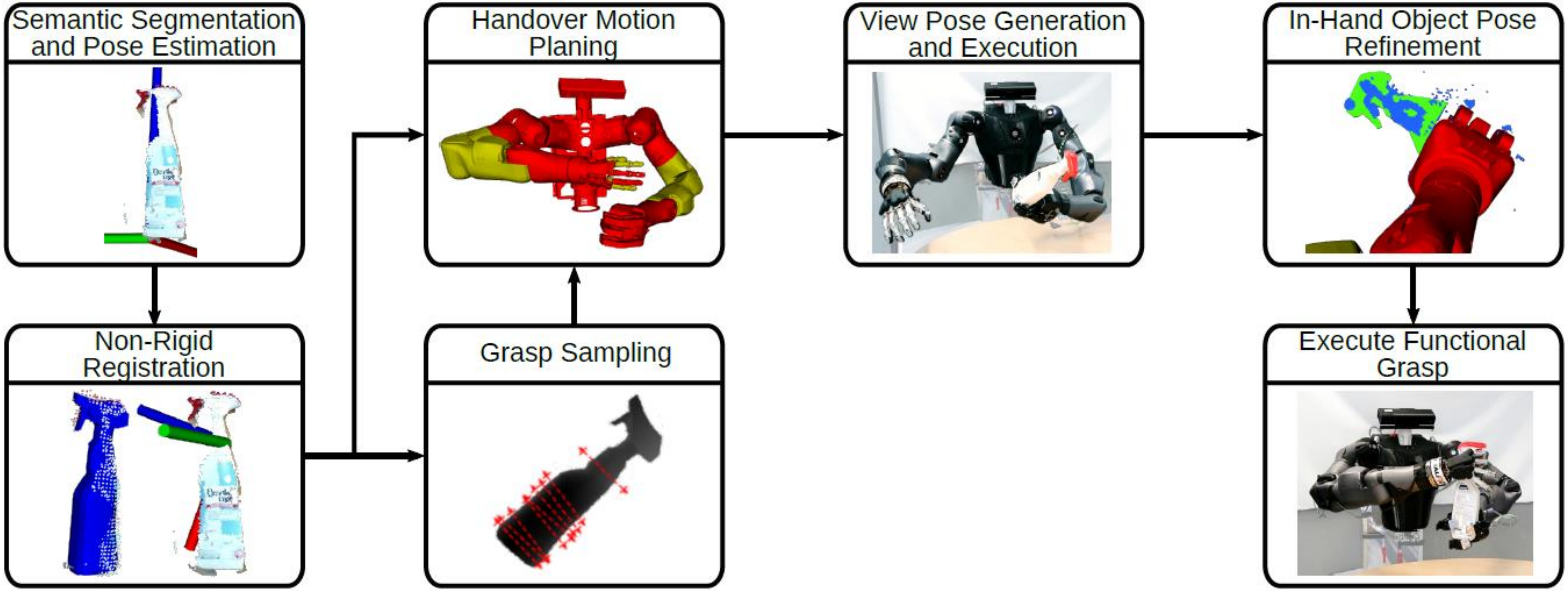
[Pavlichenko et al., IROS 2017]

Grasping an Unknown Power Drill and Fastening Screws

UNIVERSITÄT BONN AIS

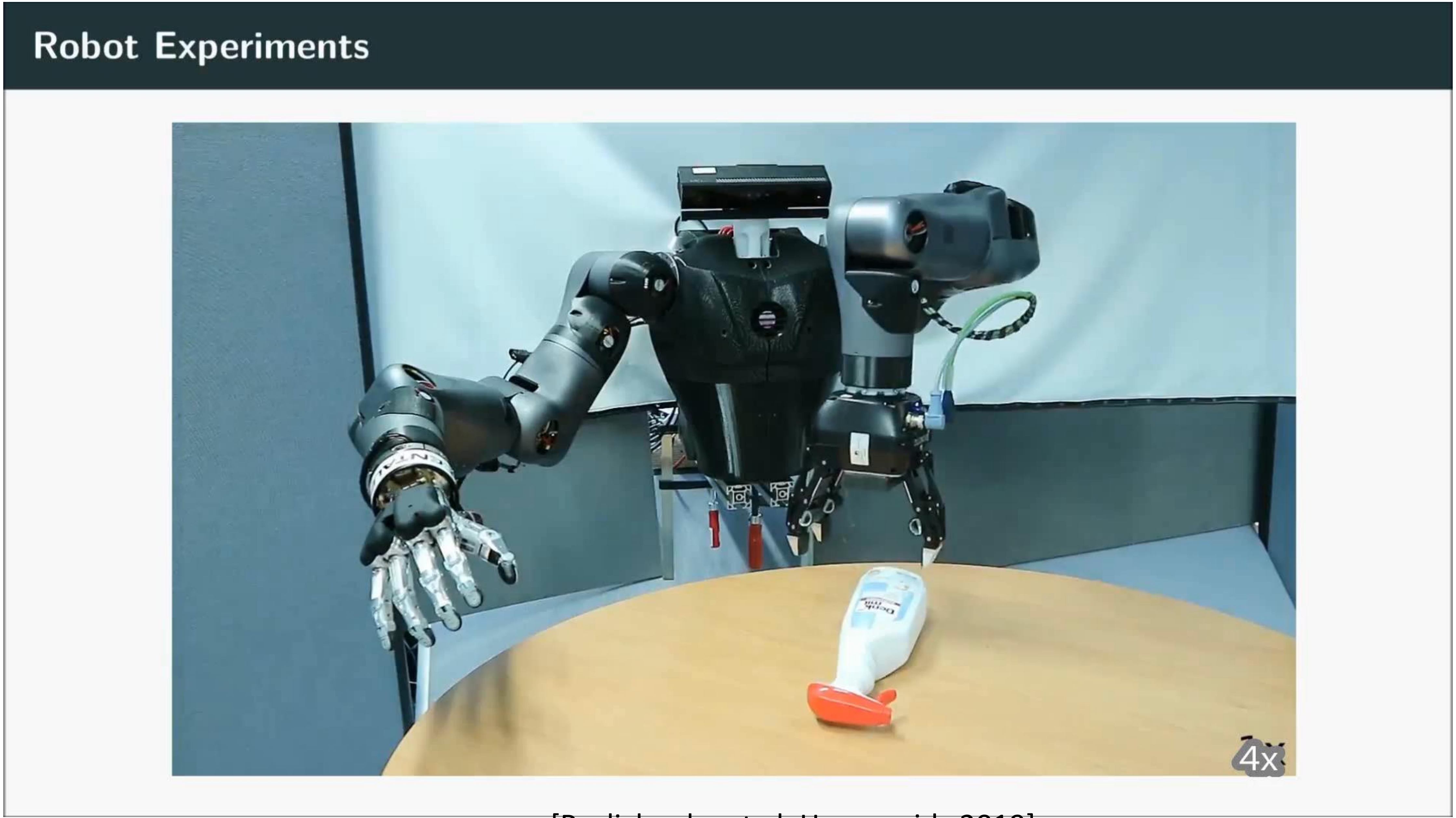
Regrasping

- Direct functional grasps not always feasible
- Pick up object with support hand, such that it can be grasped in a functional way



[Pavlichenko et al. Humanoids 2019]

Regrasping



[Pavlichenko et al. Humanoids 2019]

Conclusions

- Contributions in individual modules, integration is work-in-progress
- Structured models (e.g. rigid/non-rigid meshes) are advantageous for scene parsing
- Synthetic training data can replace real data
- Next:
  Interactive Perception!