Perception, Planning, and Learning for Cognitive Service Robots

Sven Behnke

University of Bonn
Computer Science Institute VI
Autonomous Intelligent Systems

Our Cognitive Service Robots

- Domestic service tasks in RoboCup@Home

Dynamaid

Cosero

TIAGo++

- Mobile manipulation for the support of rescue workers, telepresence

Momaro

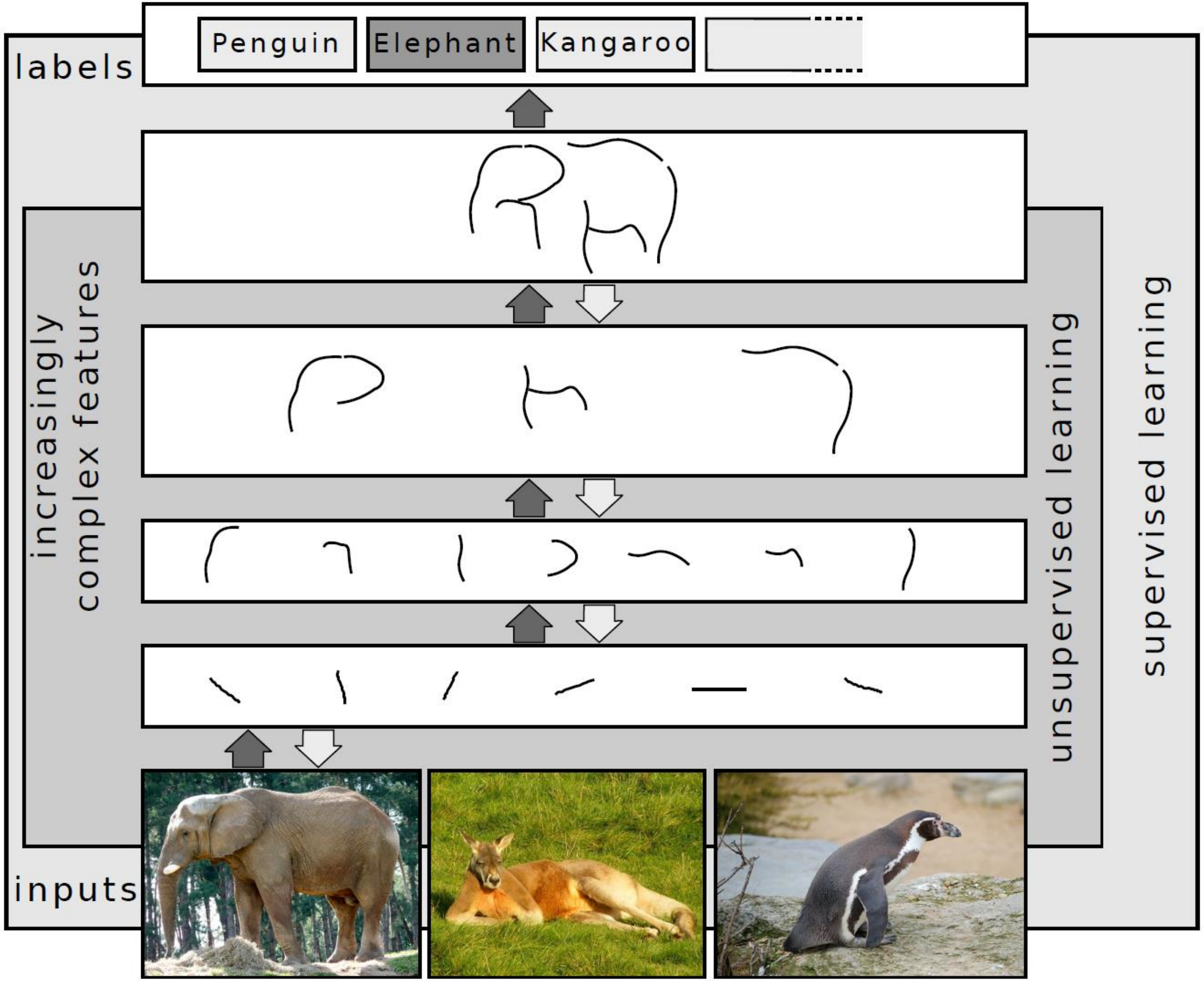
Centauro

Avatar

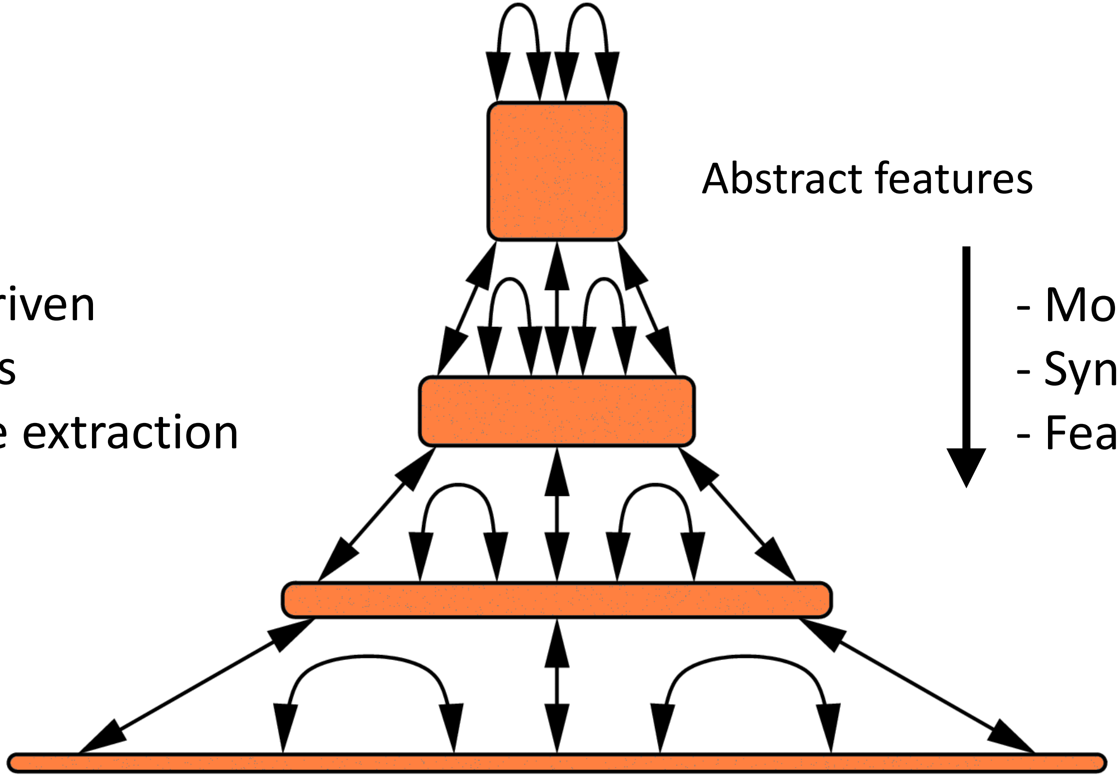
Deep Learning

- Learning layered represen-tations
- Compositionality

[Schulz; Behnke, KI 2012]



Neural Abstraction Pyramid



- Data-driven
- Analysis
- Feature extraction

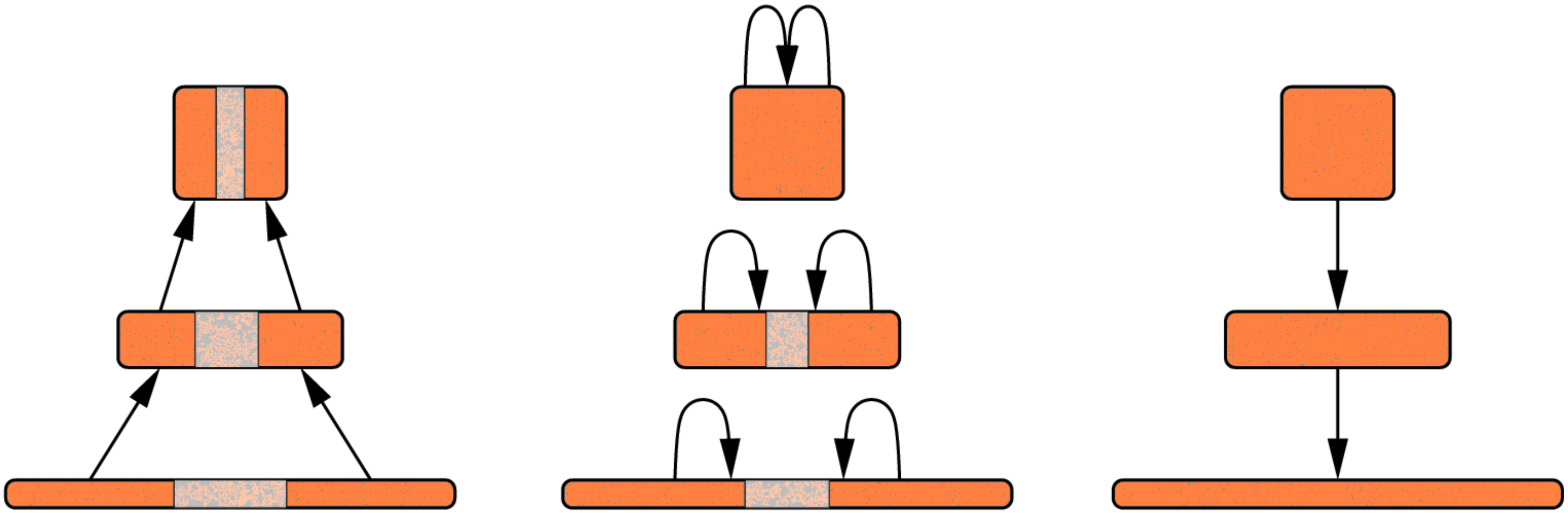
- Model-driven
- Synthesis
- Feature expansion

- Grouping - Competition - Completion

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

Iterative Image Interpretation

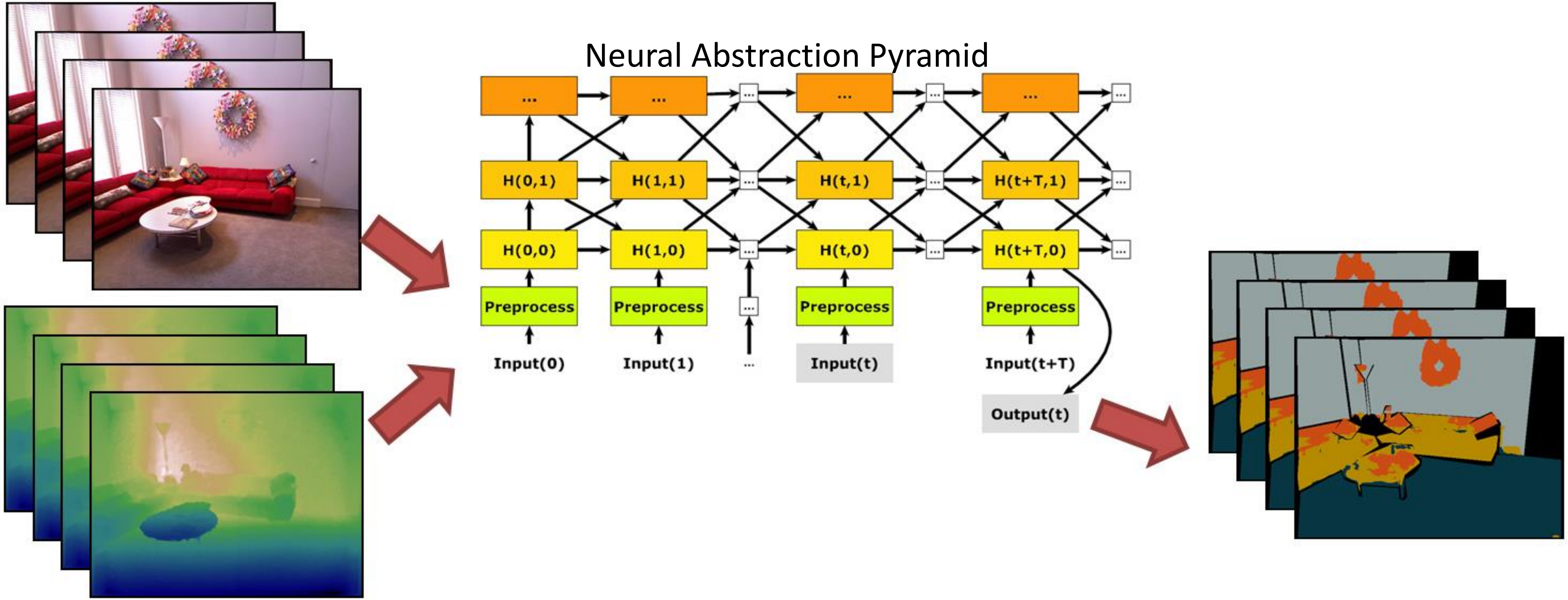
- Interpret most obvious parts first
- Use partial interpretation as context to iteratively resolve local ambiguities



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

Neural Abstraction Pyramid for Semantic Segmentation of RGB-D Video

- Recursive computation is efficient for temporal integration



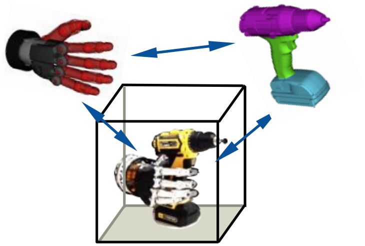
[Pavel, Schulz, Behnke, Neural Networks 2017]

The Data Problem

- Deep Learning in robotics (still) suffers from shortage of available examples
- We address this problem in three ways:

1. **Transfer learning:**
   Pre-training on large related data,
   self-supervised learning

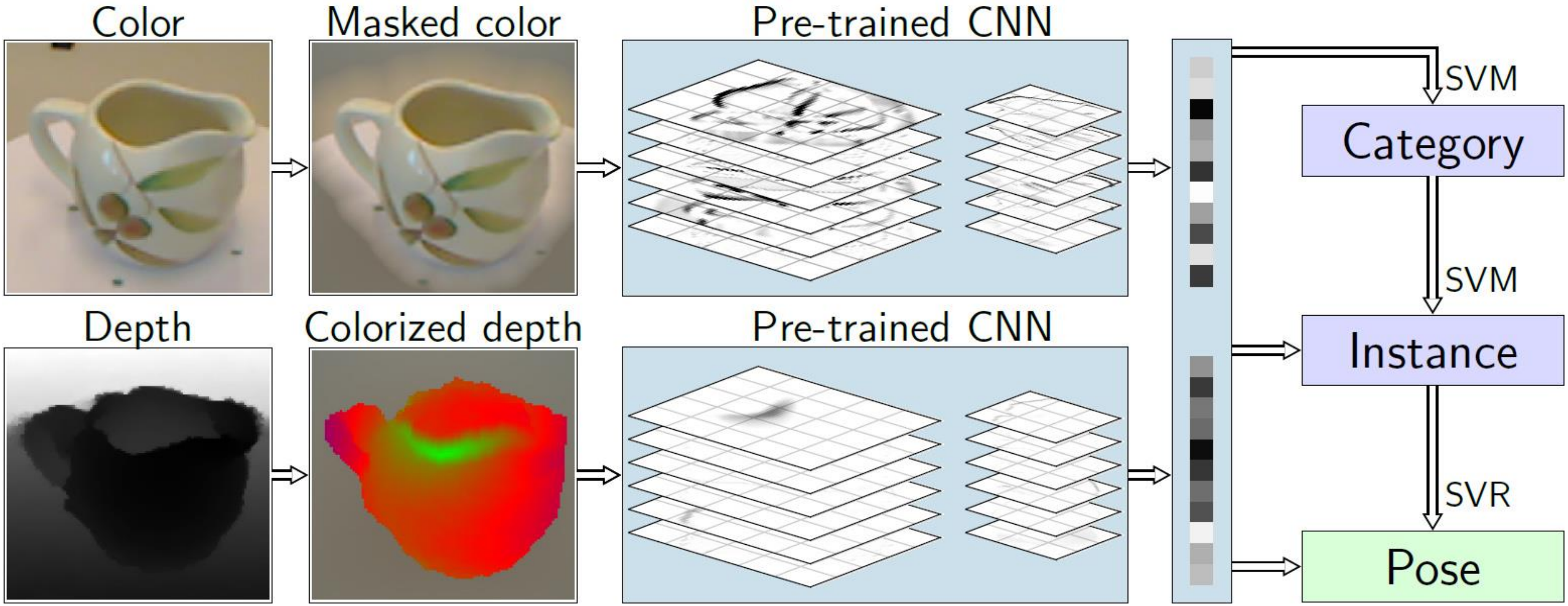
2. **Generating data**:
   Online mesh databases,
   scene synthesis



3. **Inductive biases**:
   3D projective geometry,
   camera motion, canonical frames,
   object relations, compositionality, …

RGB-D Object Recognition and Pose Estimation

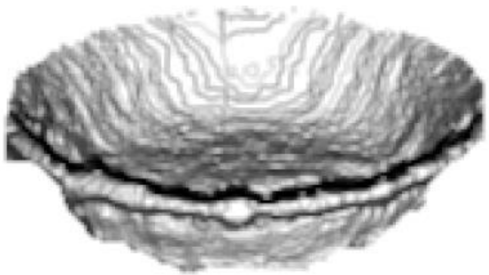
- Transfer learning from large-scale data sets

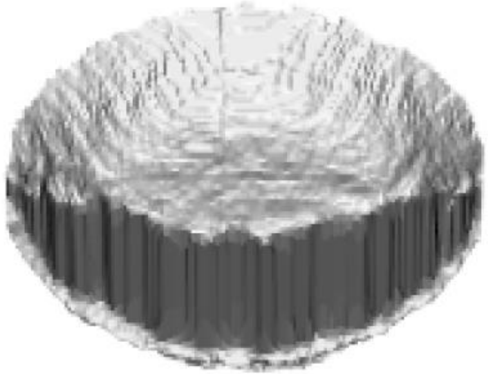


[Schwarz, Schulz, Behnke, ICRA2015]

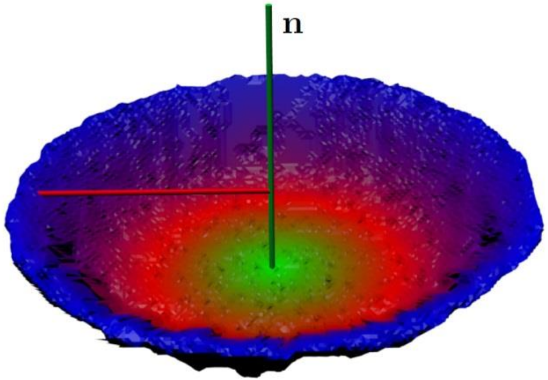
Canonical View, Colorization

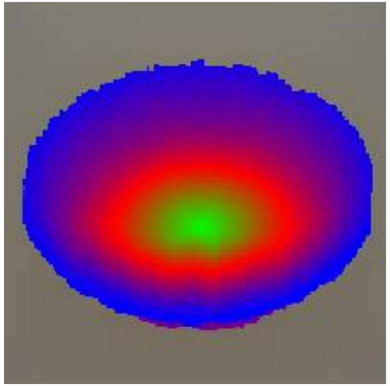
- Objects viewed from different elevation
- Render canonical view





- Colorization based on distance from center vertical

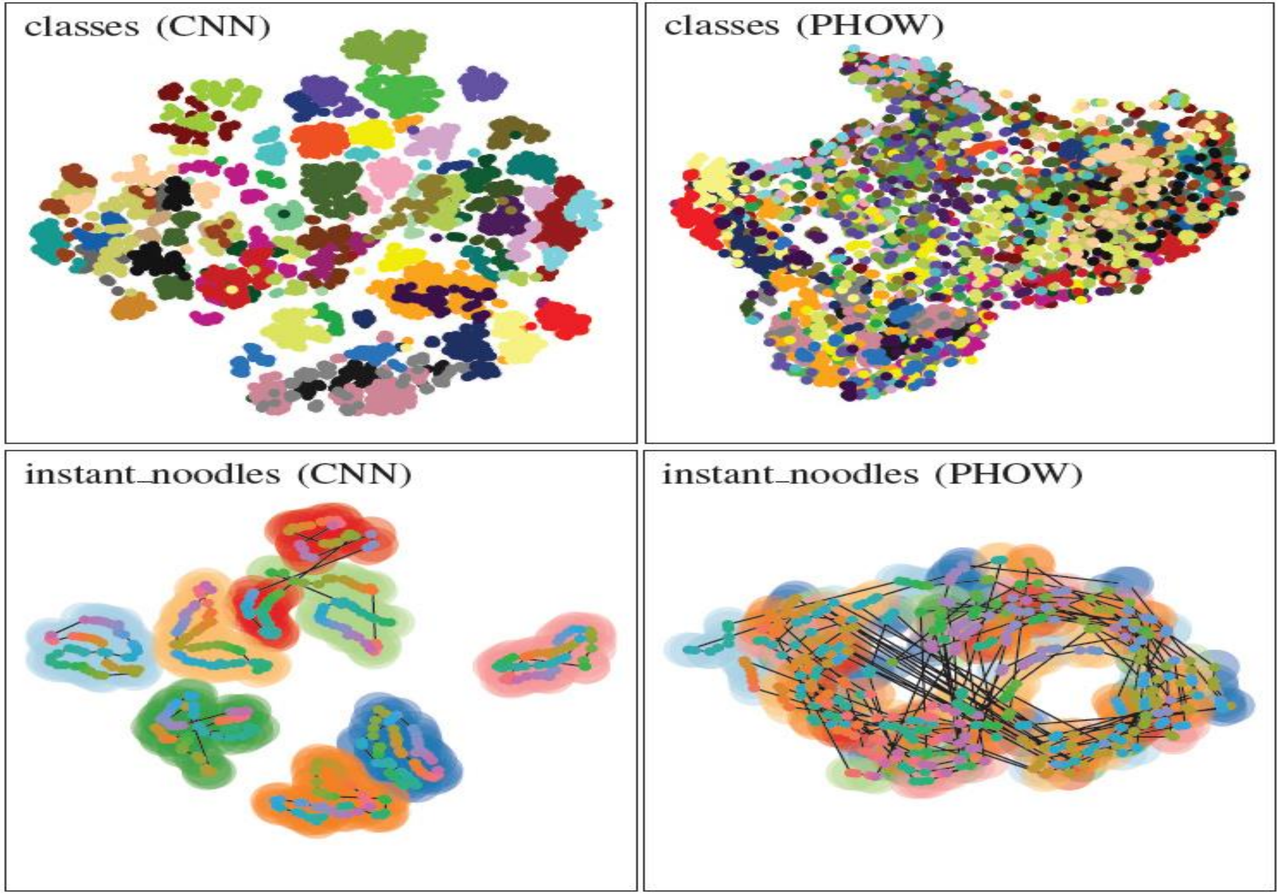




[Schwarz, Schulz, Behnke, ICRA2015]

Pretrained Features Disentangle Data

- t-SNE embedding



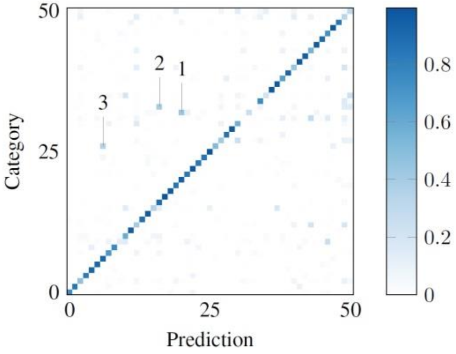
[Schwarz, Schulz, Behnke ICRA2015]

Recognition Accuracy

- Improved both category and instance recognition

	Category Accuracy (%)		Instance Accuracy (%)	
Method	RGB	RGB-D	RGB	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:



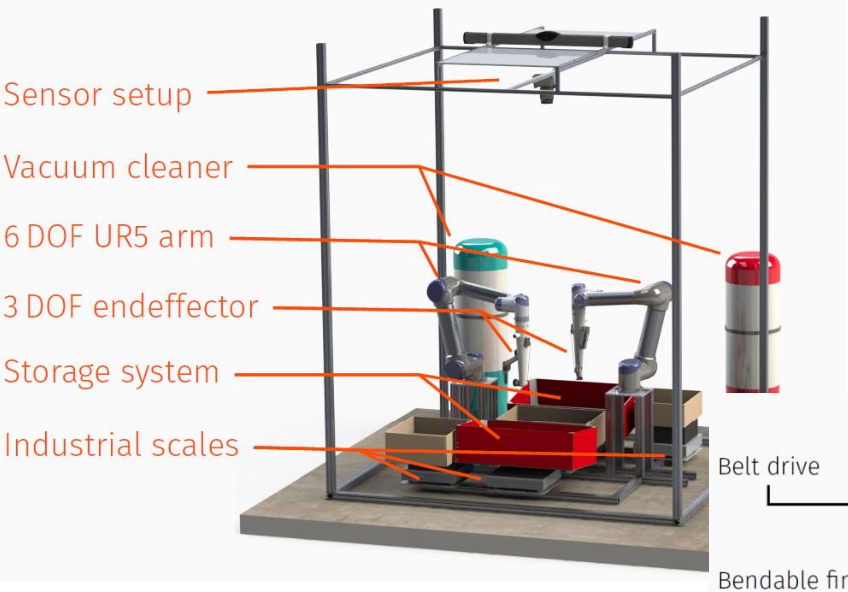
1: pitcher / coffe mug

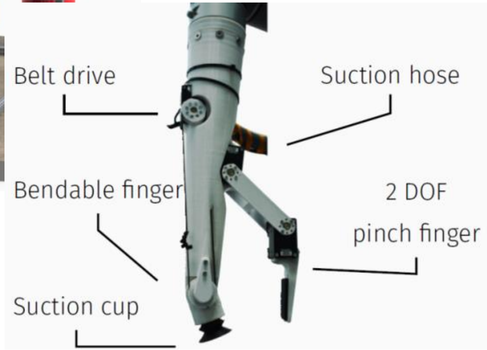
2: peach / sponge

[Schwarz, Schulz, Behnke, ICRA2015]

Amazon Robotics Challenge

- Storing and picking of items
- Dual-arm robotic system





[Schwarz et al. ICRA 2018]

[Amazon]

Object Capture and Scene Rendering

- Turntable + DLSR camera

- Insertion in complex annotated scenes

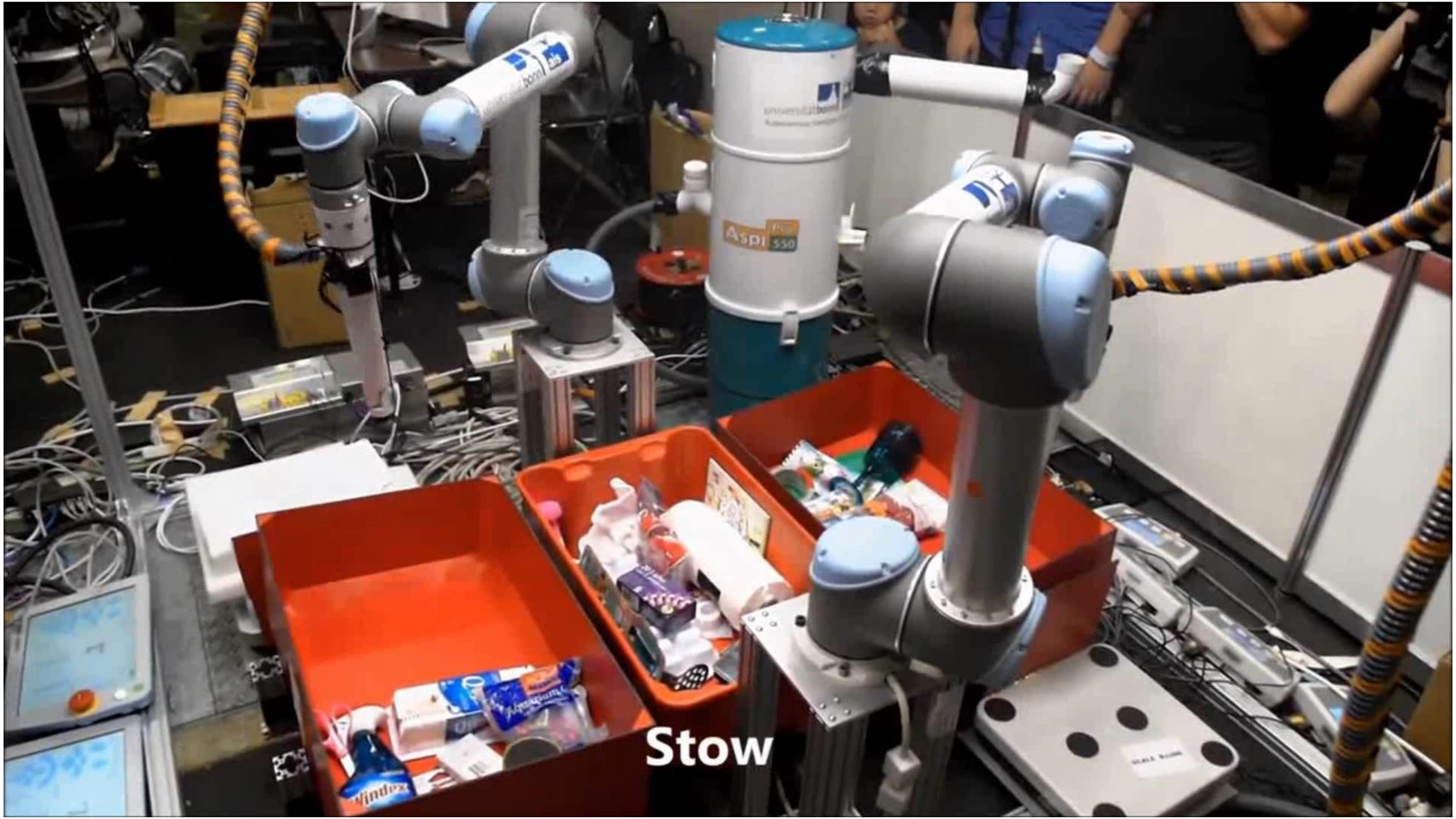
[Schwarz et al. ICRA 2018]

Semantic Segmentation and Grasp Pose Estimation

- Semantic segmentation using RefineNet [Lin et al. CVPR 2017]
- Grasp positions in segment centers

[Schwarz et al. ICRA 2018]

Amazon Robotics Challenge 2017

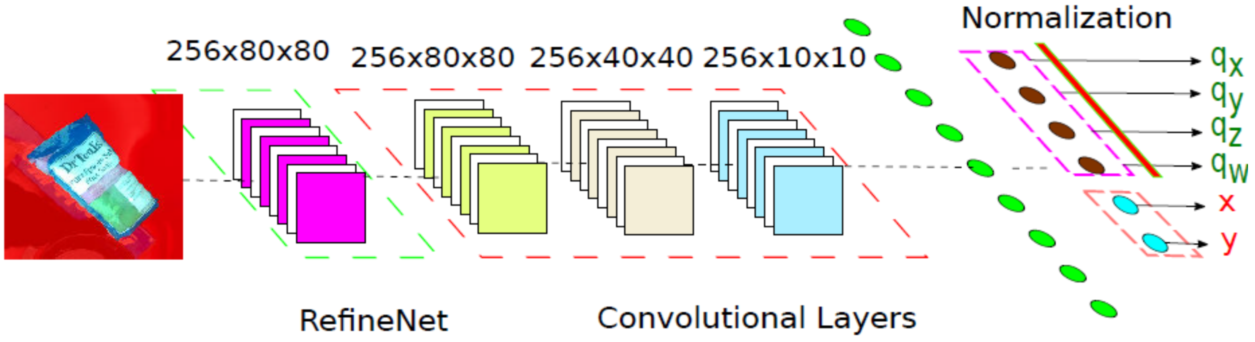


[Schwarz et al. ICRA 2018]

UNIVERSITÄT BONN AIS

Object Pose Estimation

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



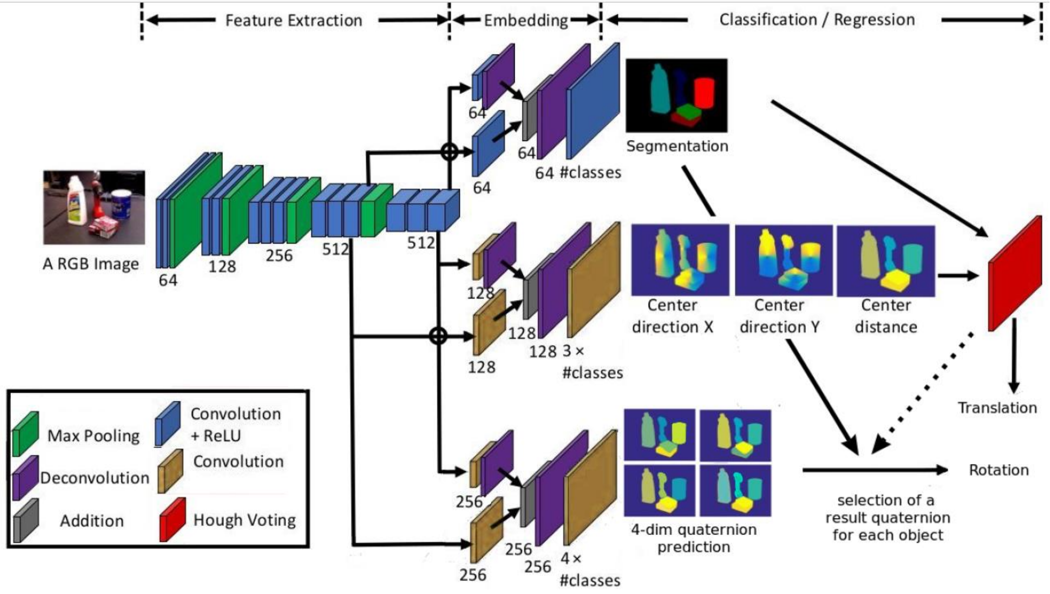
Input

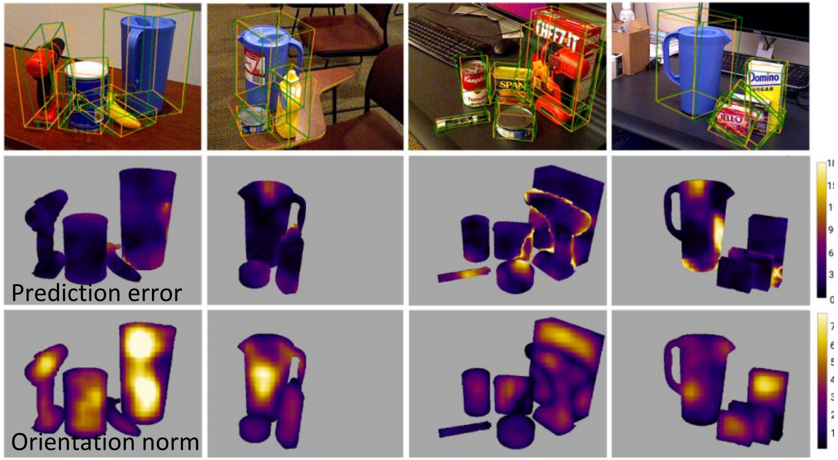
Predicted pose

[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 center and orientation, without cutting out

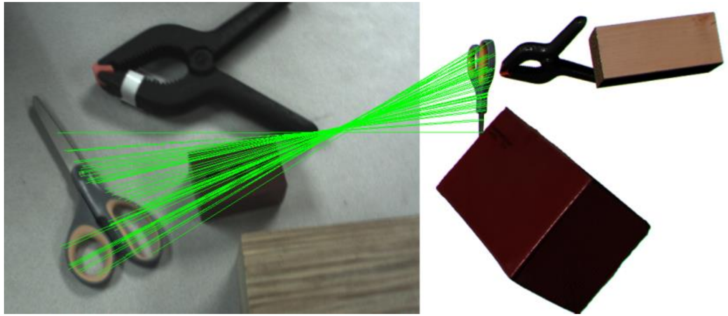




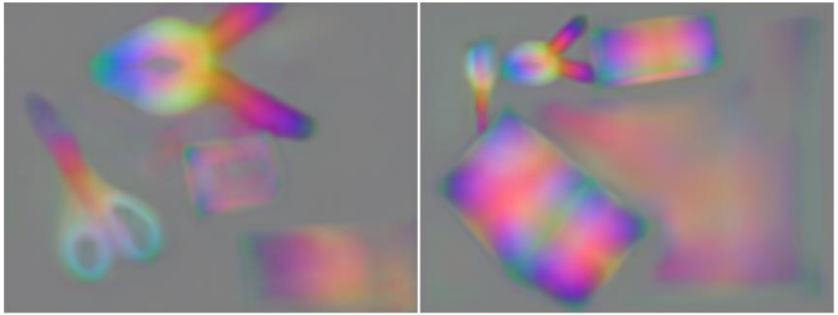
[Capellen et al., VISAPP 2020]

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



Known correspondences



Learned features

[Periyasamy, Schwarz, Behnke Humanoids 2019]

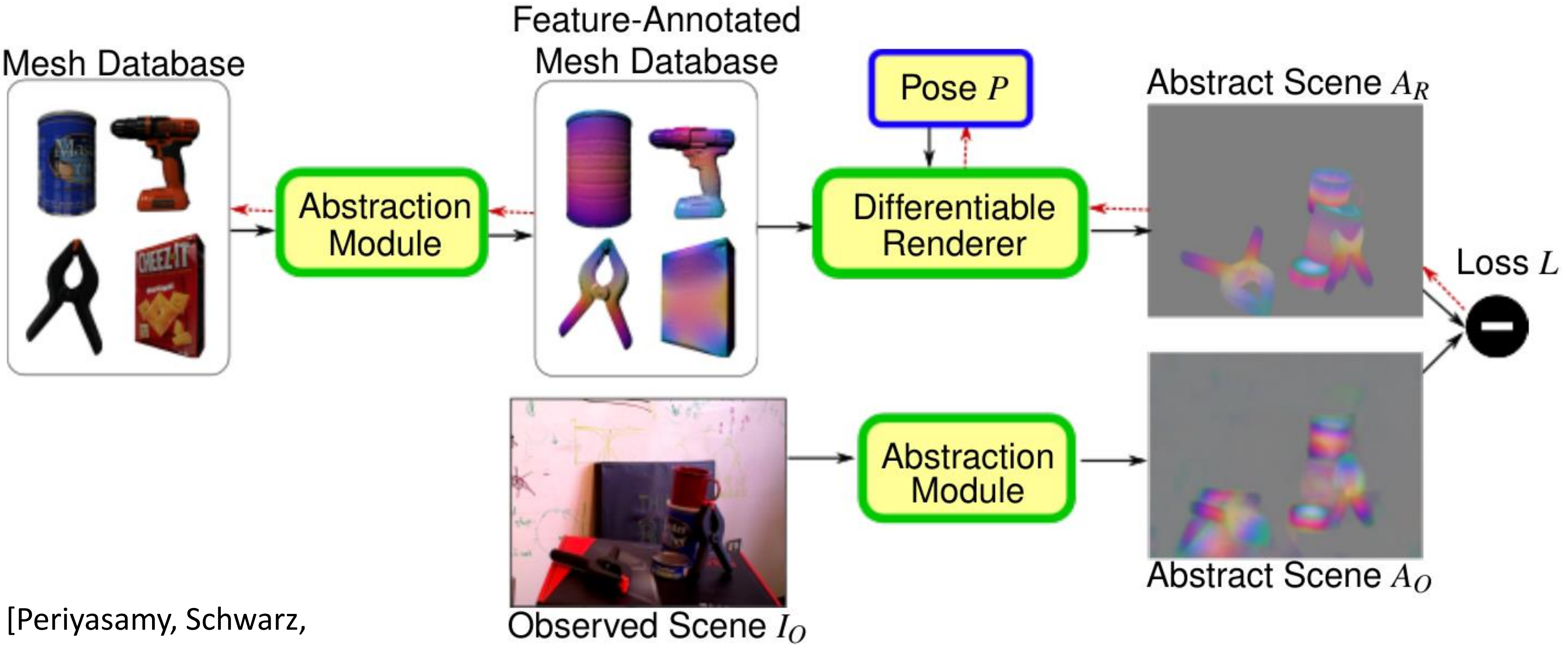
Descriptors as Texture on Object Surfaces

- 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



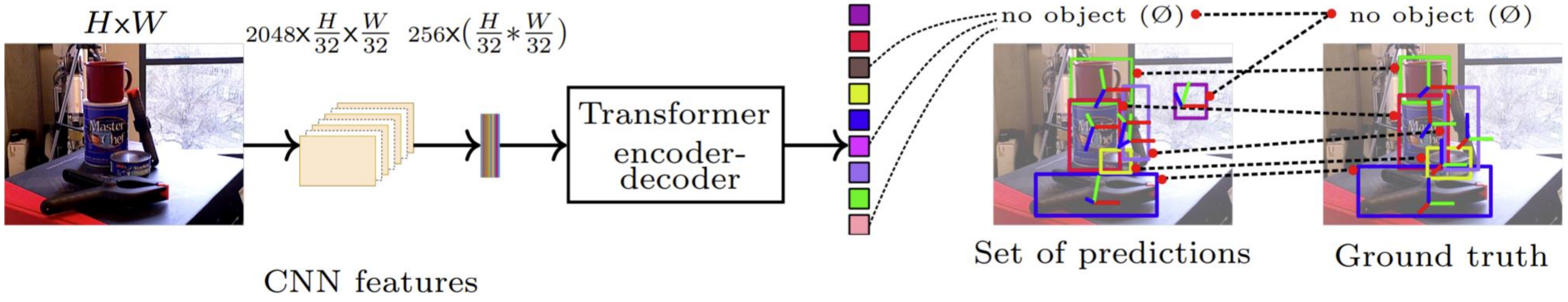
[Periyasamy, Schwarz, Behnke Humanoids 2019]

Registration Examples

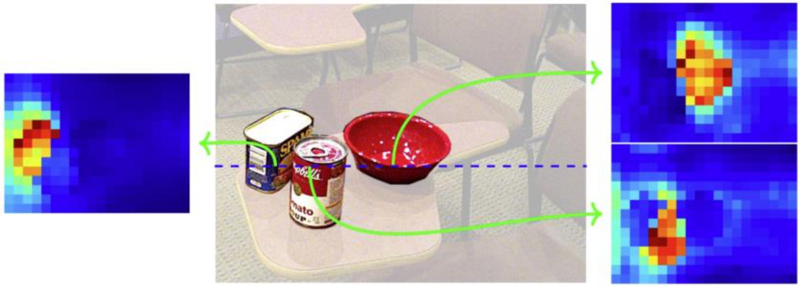
[Periyasamy, Schwarz, Behnke Humanoids 2019]

T6D-Direct: Transformers for Multi-Object 6D Pose Direct Regression

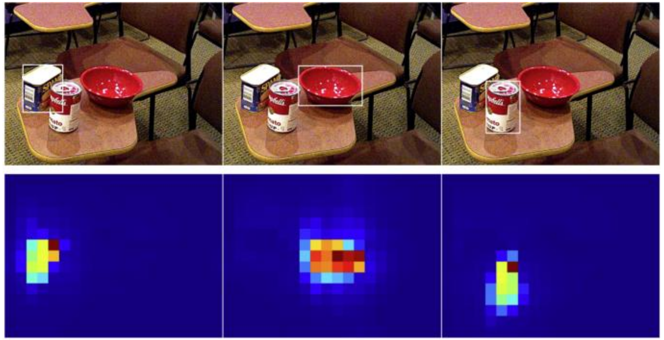
- Extends DETR: End-to-end object detection with transformers [Carion et al. ECCV 2020]
- End-to-end differentiable pipeline for 6D object pose estimation



Encoder self-attention

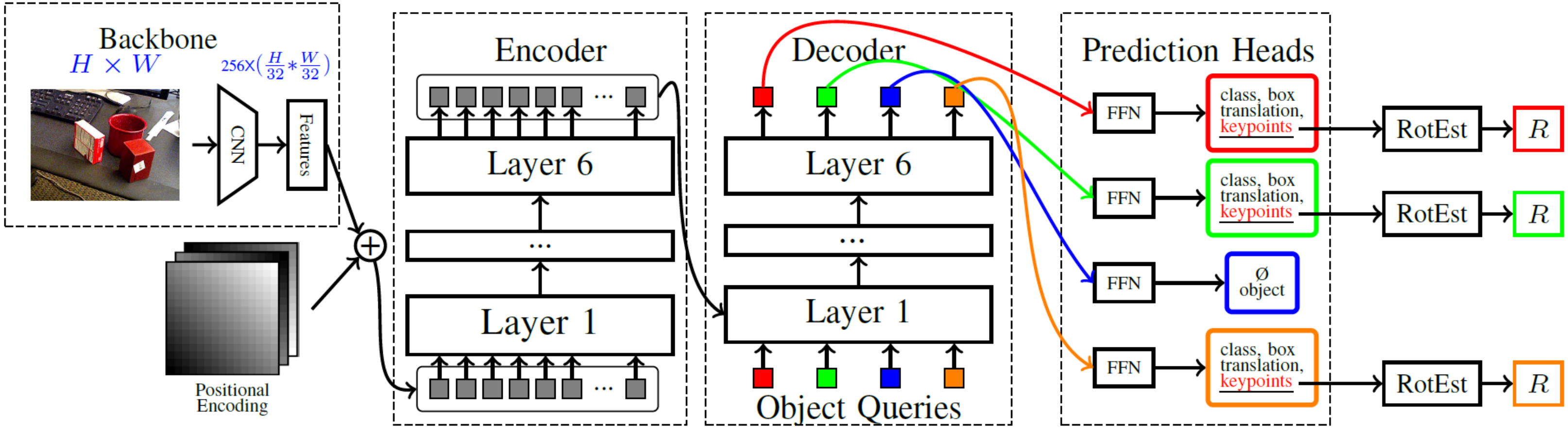


Object detections and decoder attention



[Amini et al. GCPR 2021]

YOLOPose: Multi-Object 6D Pose Estimation using Keypoint Regression

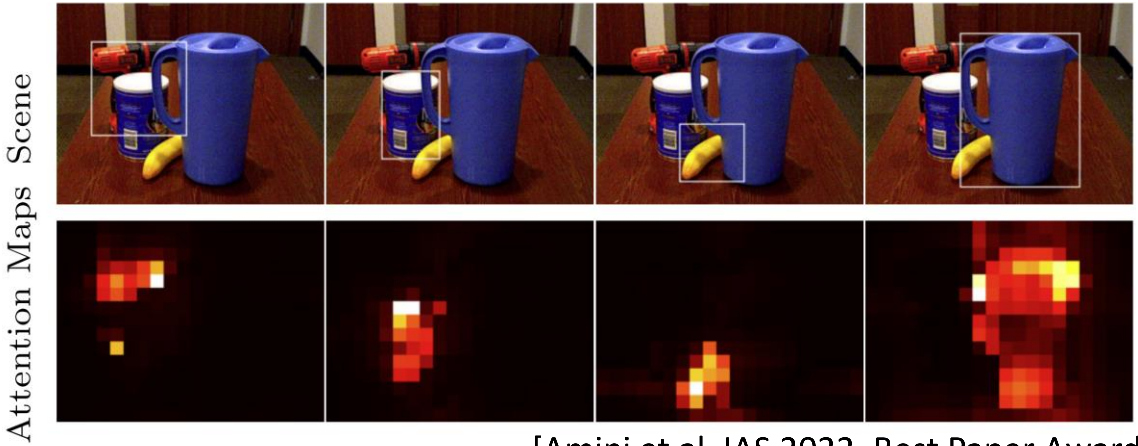


[Amini et al. IAS 2022, Best Paper Award]

YOLOPose: Multi-Object 6D Pose Estimation using Keypoint Regression

- Encoder self-attention

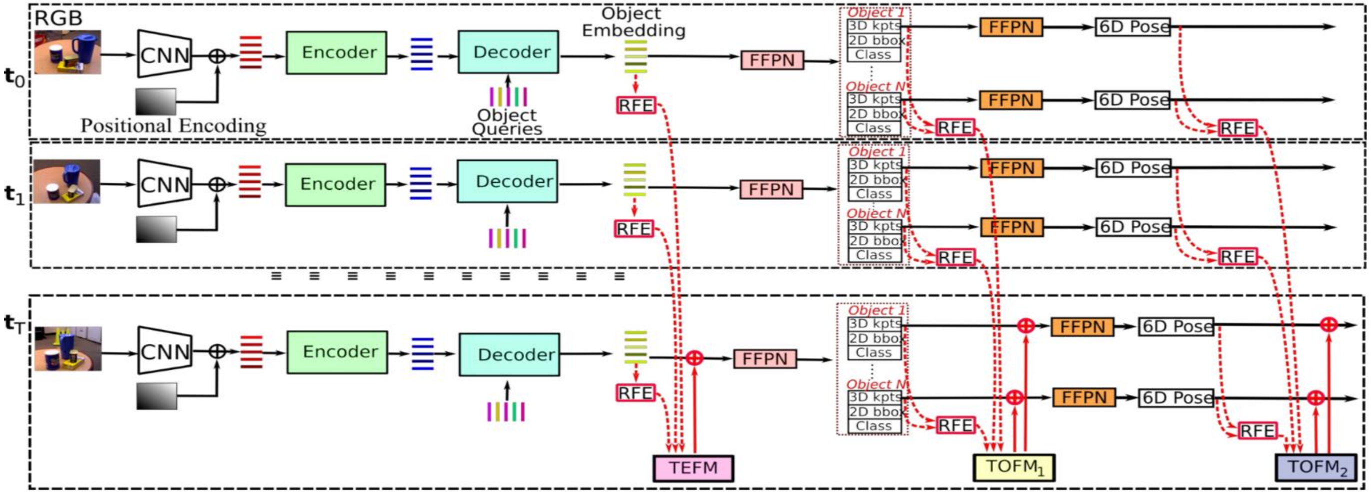
- Object detections and decoder cross-attention

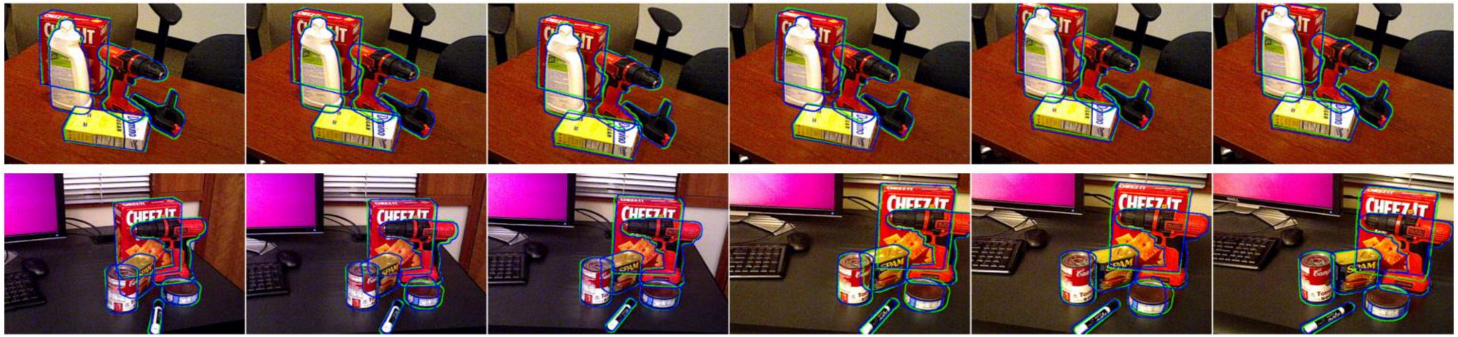


[Amini et al. IAS 2022, Best Paper Award]

MOTPose: Attention-based Temporal Fusion for Multi-object 6D Pose Estimation

- Propagating object embeddings, object descriptors, and poses

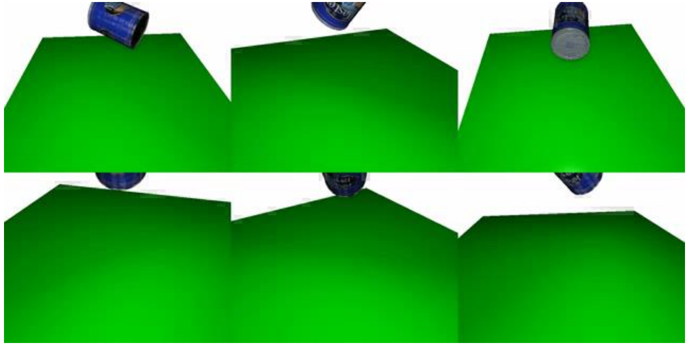


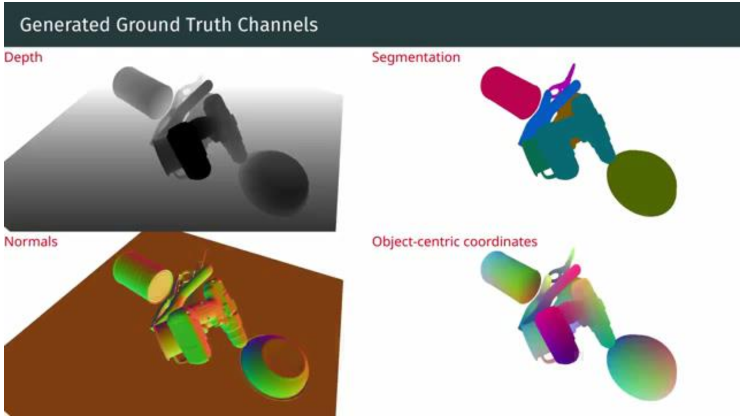


[Periyasamy, ICRA 2024]

Stillleben: 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

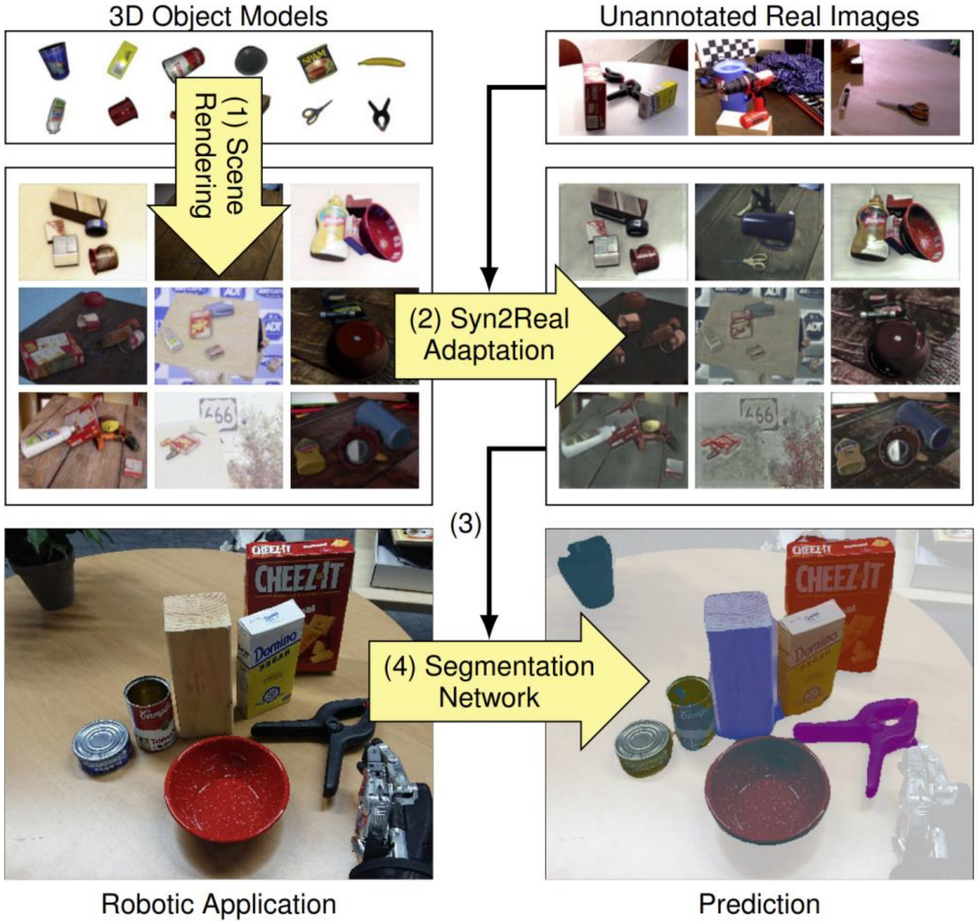




[Schwarz and Behnke, ICRA 2020]

Synthetic-to-Real Domain Adaptation

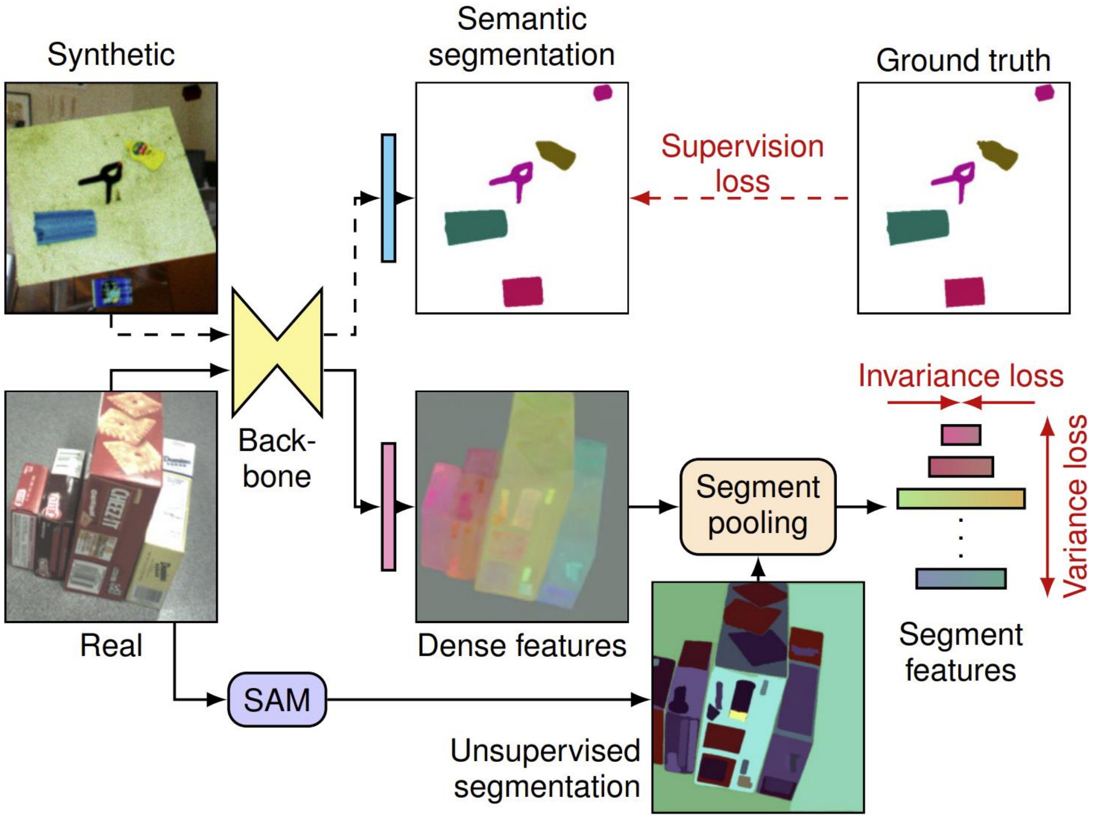
- Generate images from 3D object meshes
- Adapt the synthetic images to the real domain using un-annotated real images (GAN loss)
- Train downstream task using adapted images
- Semantic segmentation results almost as good as trained with real images
- Improved results in combina-tion with real annotations



[Imbusch et al. CASE 2022]

Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

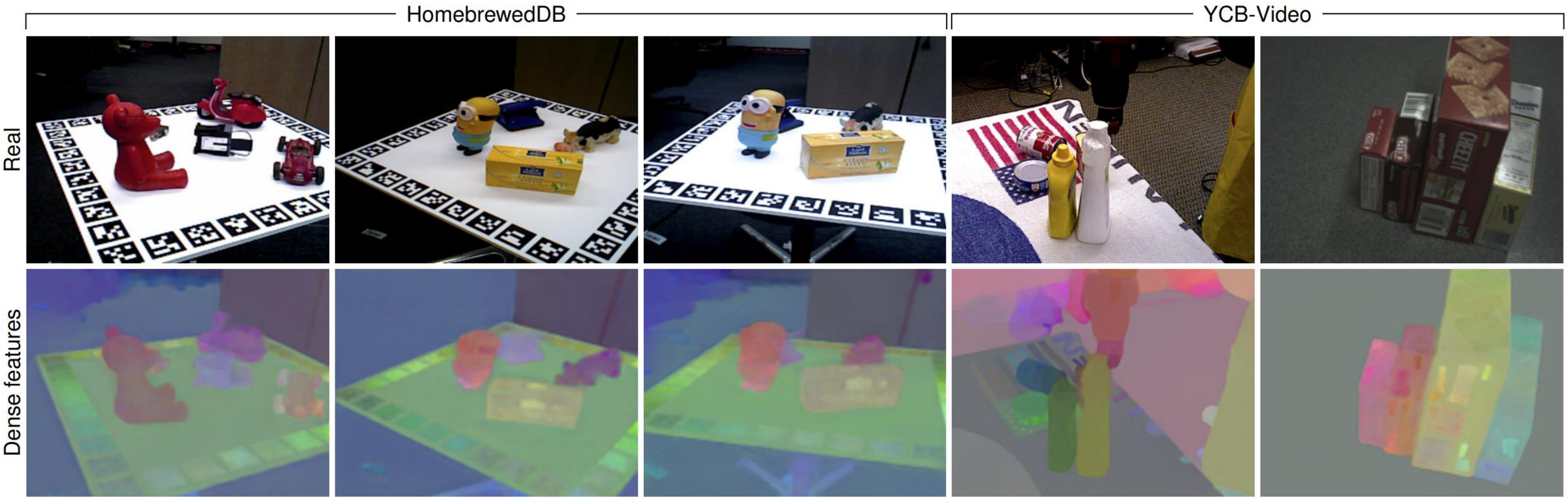
- Learns from synthetic scenes and unannotated real images
- Supervised training of semantic segmentation for synthetic scenes
- Segment Anything Model (SAM) used to generate many overlapping segments for real images
- Dense features from shared backbone
- Contrastive loss for segments
  - Features within a segment are trained to have low variance
  - Features for different segments trained to have high variance



[Bonani et al. arXiv:2309.15562]

Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

- Learned dense features correspond well to objects, are stable under camera motion, and label sub-parts

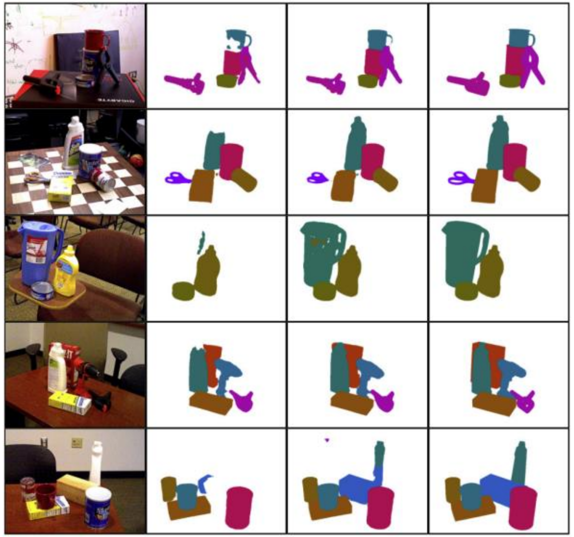


[Bonani et al. arXiv:2309.15562]

Learning from SAM: Sim2Real Domain Adaptation through Segment VI-Regularization

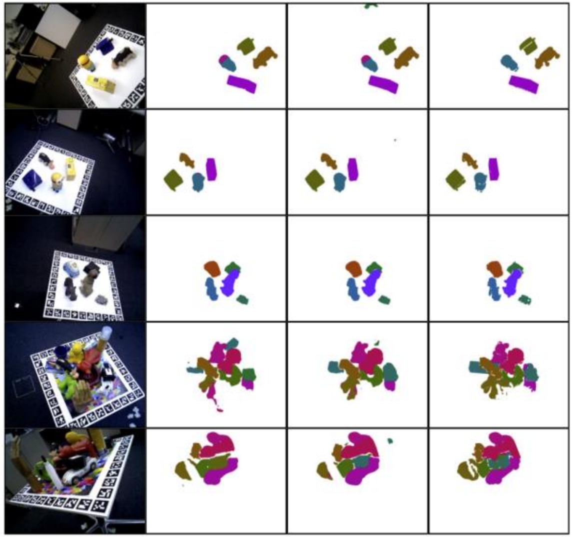
- Good results on real images without need for real labels
- Better than training with real labels on VCB-Video

YCB-Video



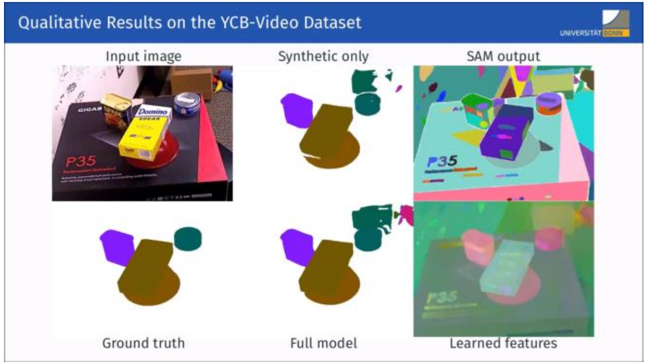
Input image a) Synthetic-only b) Full model Ground truth

HomebrewedDB



Input image a) Synthetic-only b) Full model Ground truth

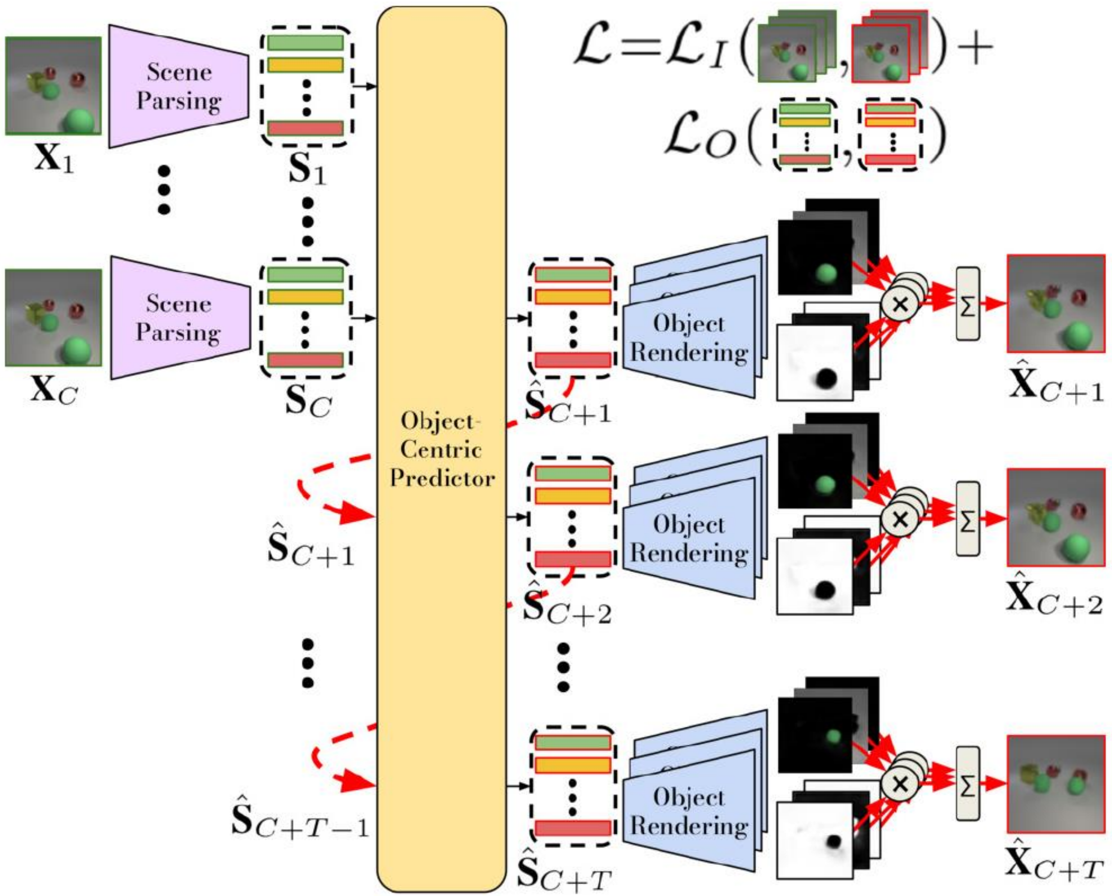
[Bonani et al. arXiv:2309.15562]



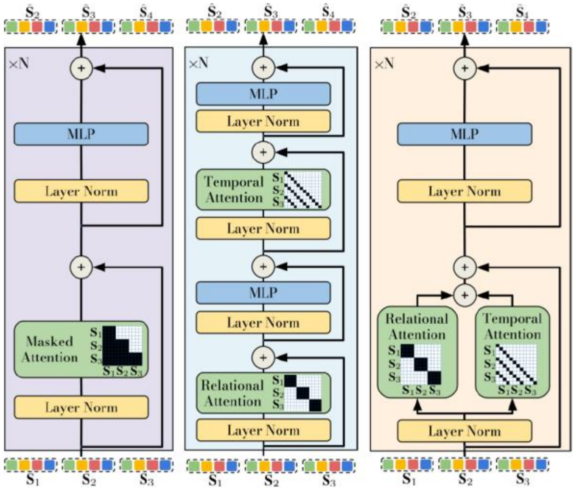
Method	Mean IoU YCB-Video [19]	HomebrewedDB [20]
Imbusch *et al.* [4]		
- real labels	0.770	0.737
- synthetic only	0.701	0.481[1]
- full	0.763	0.558[1]
Ours		
- real labels	0.839	**0.883**
- synthetic only	0.807	0.748
- CUT [4] only[2]	0.814	-
- full	**0.853**	0.787[3]

Note: "real labels" is a baseline which has access to real supervision.
[1] Using Stillleben [2] synthetic data, where we use BlenderProc4BOP.
[2] Training our backbone on CUT-refined synthetic data.
[3] Model was trained for only 200k epochs.

Object-centric Video Prediction Decoupling Dynamics and Interaction



- Scene parsing into object slots
- Video synthesis from objects and masks
- Predictor decouples tempo-ral and relational attention



[Villar-Corrales et al. ICIP 2023]

Object-centric Video Prediction Data Sets

Obj3D

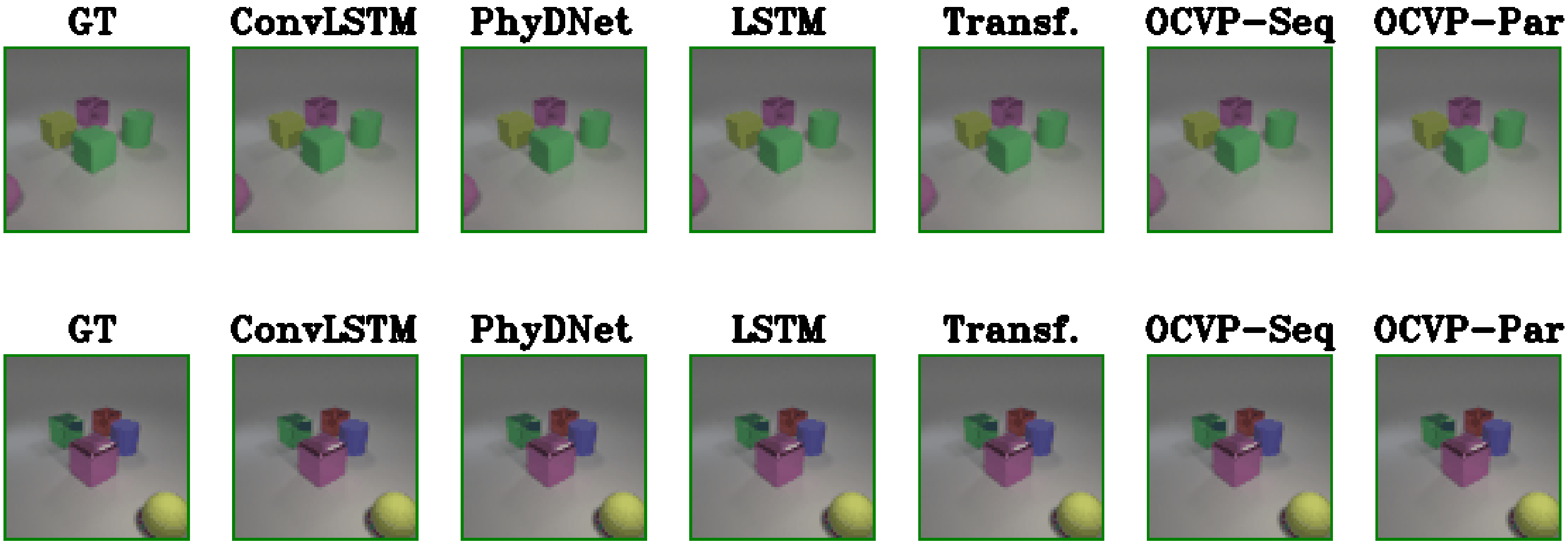
- Synthetic 3D objects
- Ball colliding with static objects
- Given 5 frames, predict next 5

MOVi-A

- Synthetic 3D objects
- Complex dynamics and occlusions
- Given 6 frames, predict next 8

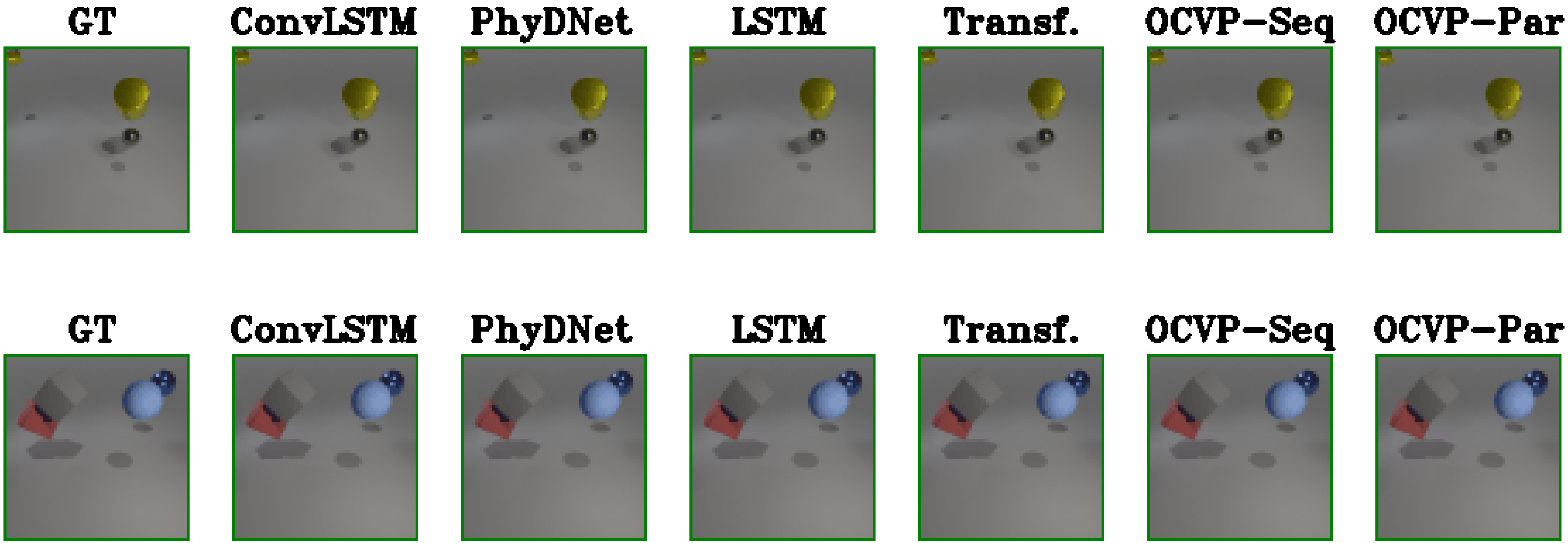
[Villar-Corrales et al. ICIP 2023]

Object-centric Video Prediction: Obj3D



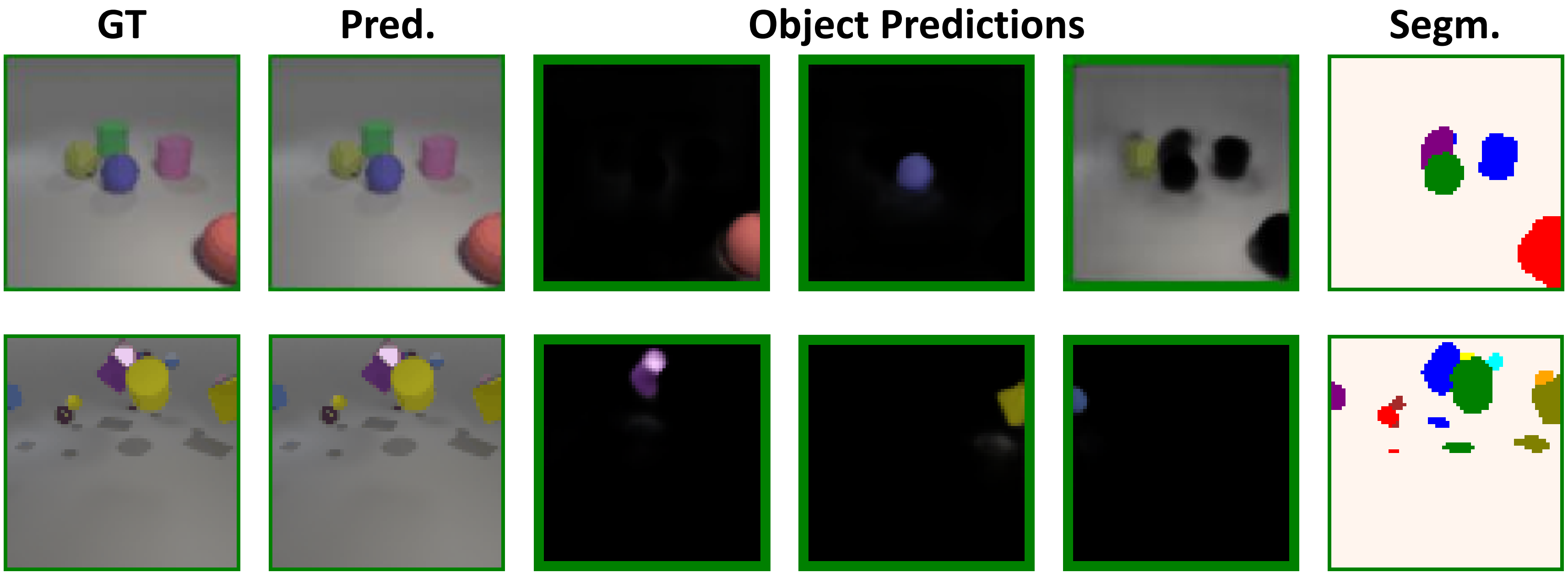
[Villar-Corrales et al. ICIP 2023]

Object-centric Video Prediction: MOVi-A



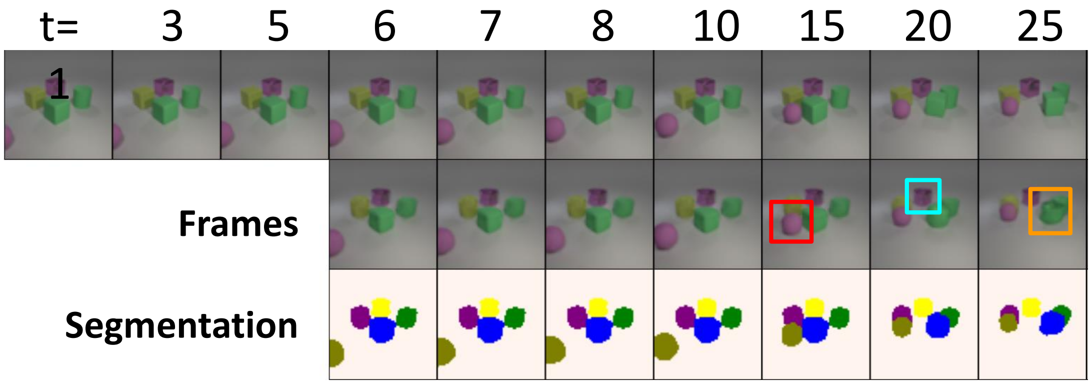
[Villar-Corrales et al. ICIP 2023]

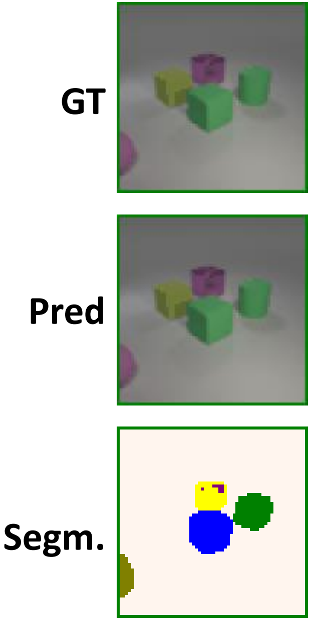
Object-centric Video Prediction: Object Predictions

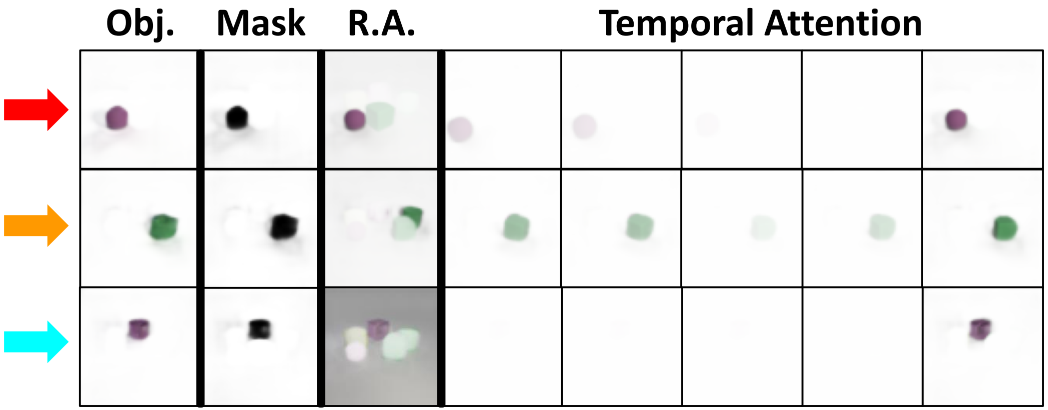


[Villar-Corrales et al. ICIP 2023]

Object-centric Video Prediction: Model Interpretability



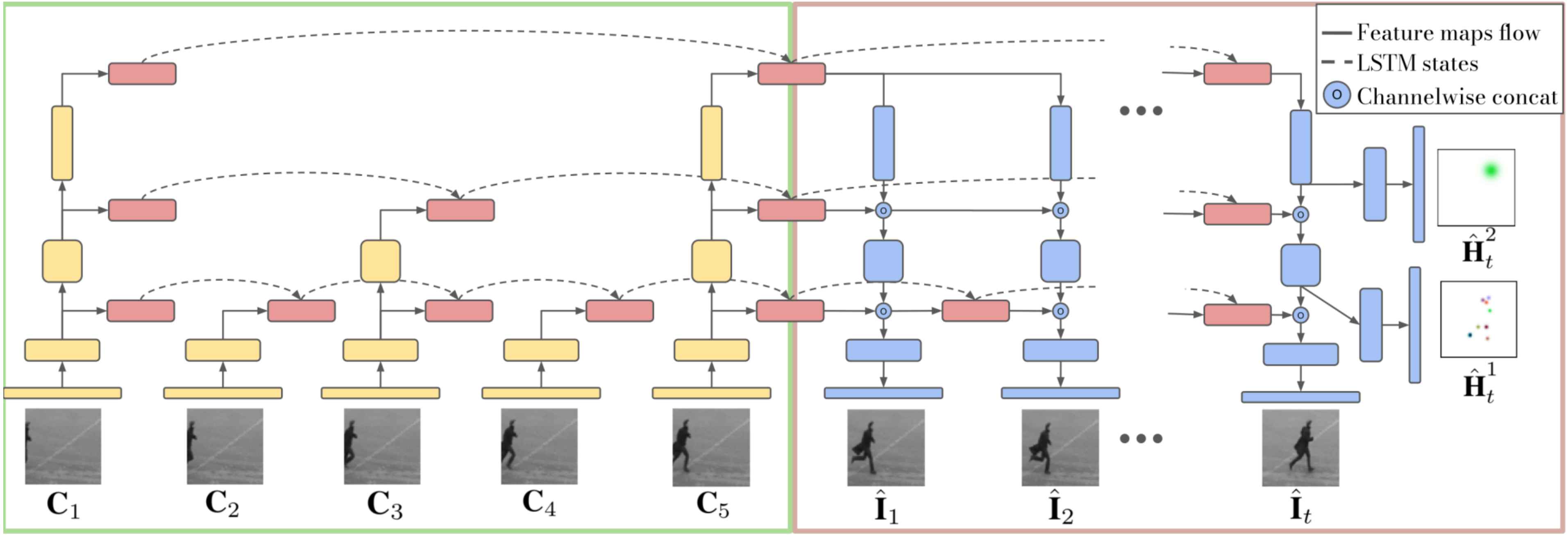




[Villar-Corrales et al. ICIP 2023]

MSPred: Video Prediction at Multiple Spatio-Temporal Scales

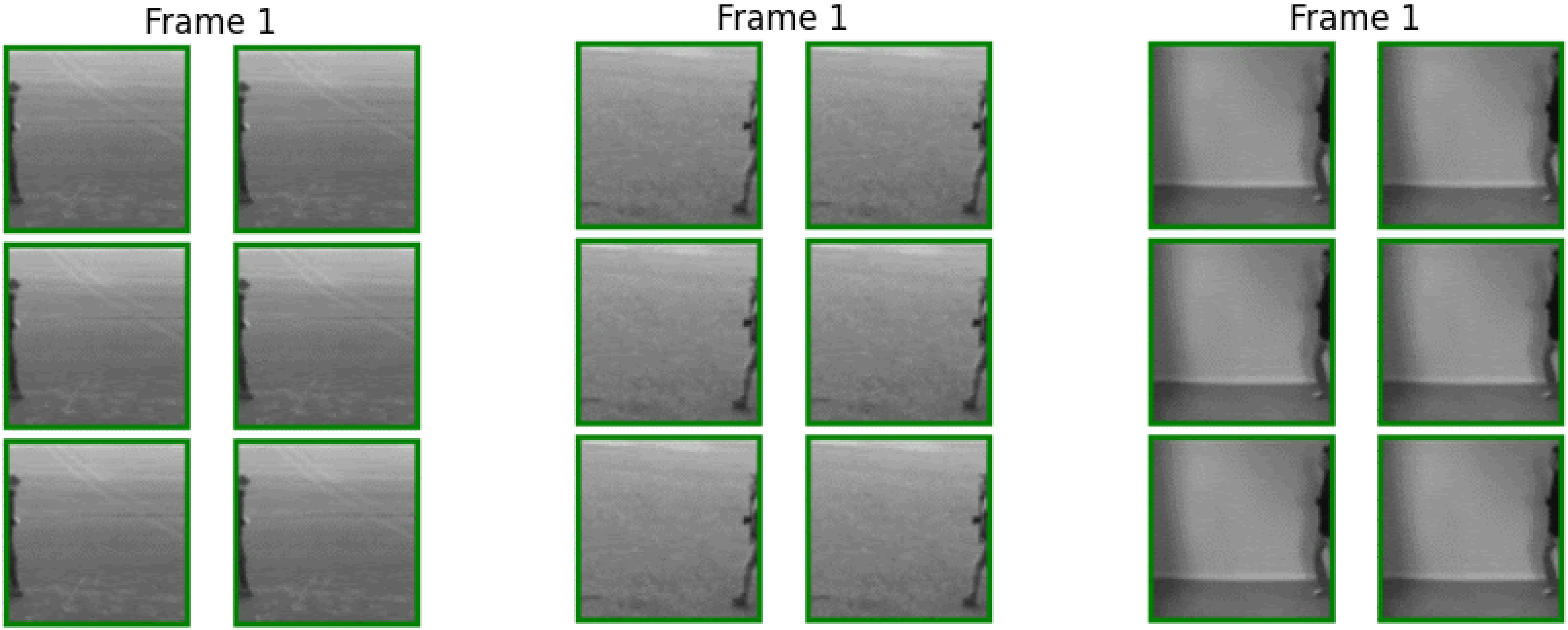
- Coarser, more abstract predictions for longer time horizons in higher layers
- Predict image itself, human pose joint keypoints, and human body position



[Villar-Corrales et al., BMVC 2022]

MSPred: Video Prediction at Multiple Spatio-Temporal Scales

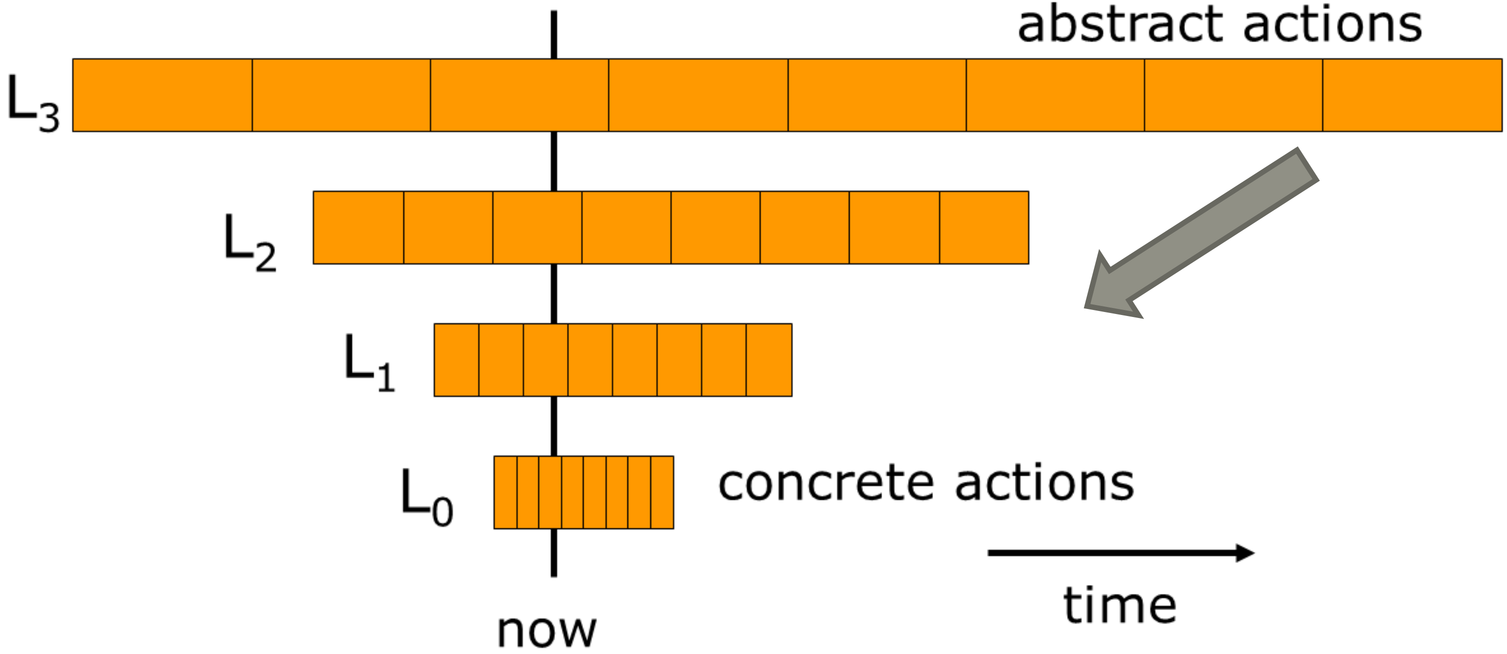
- Coarser, more abstract predictions for longer time horizons in higher layers
- Predict image itself, human pose joint keypoints, and human body position



[Villar-Corrales et al., BMVC 2022]

Hierarchical Planning in the Now

- Use predicted state on different layers of abstraction for planning
- Coarse-to-fine planning makes actions more concrete as they come closer to execution
- Plan consists of few steps on each layer



Centauro Robot

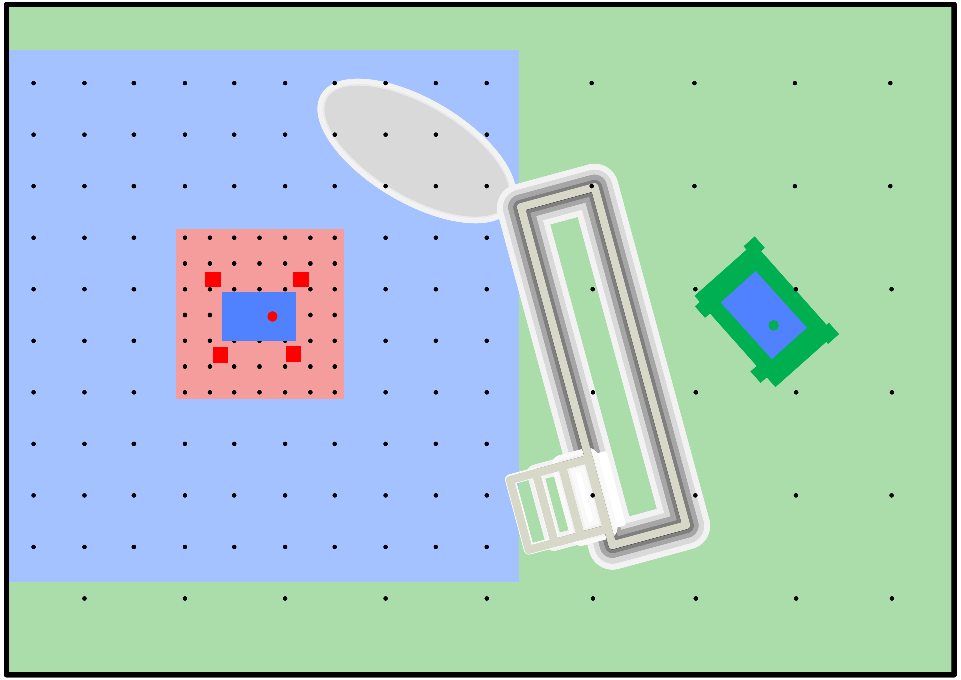
CENTAURO

- Serial elastic actuators
- 42 main DoFs
- Schunk hand
- 3D laser
- RGB-D camera
- Color cameras
- Two GPU PCs

[Tsagarakis et al., IIT 2017]

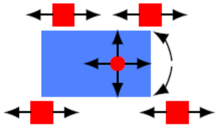
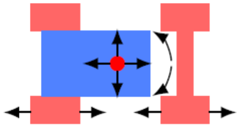
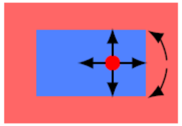
Hybrid Driving-Stepping Locomotion Planning: Abstraction

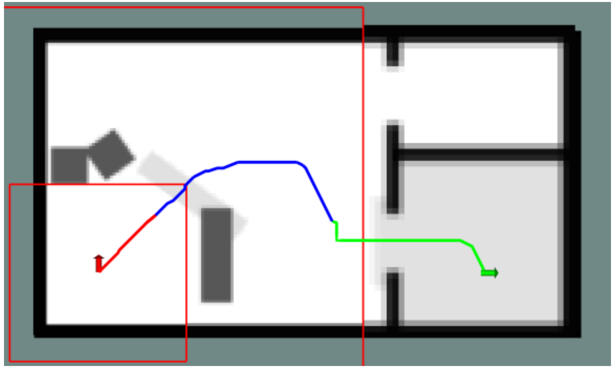
- Planning in the here and now
- Far-away details are abstracted away

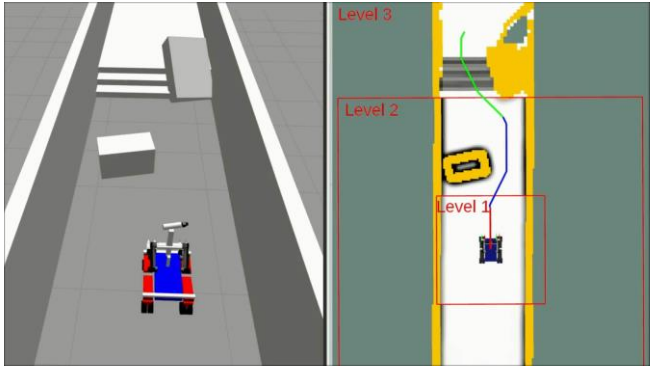


[Klamt and Behnke, IROS 2017, ICRA 2018]

Hybrid Driving-Stepping Locomotion Planning: Abstraction

Level	Map Resolution	Map Features	Robot Representation	Action Semantics
1	• 2.5 cm • 64 orient.	• Height		• Individual Foot Actions
2	• 5.0 cm • 32 orient.	• Height • Height Difference		• Foot Pair Actions
3	• 10 cm • 16 orient.	• Height • Height Difference • Terrain Class		• Whole Robot Actions

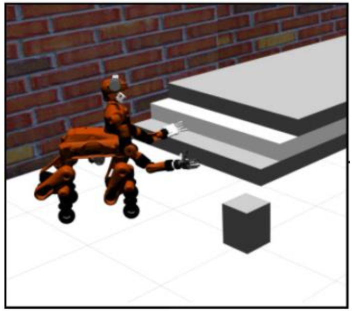




[Klamt and Behnke, IROS 2017, ICRA 2018]

Learning Cost Functions of Abstract Representations

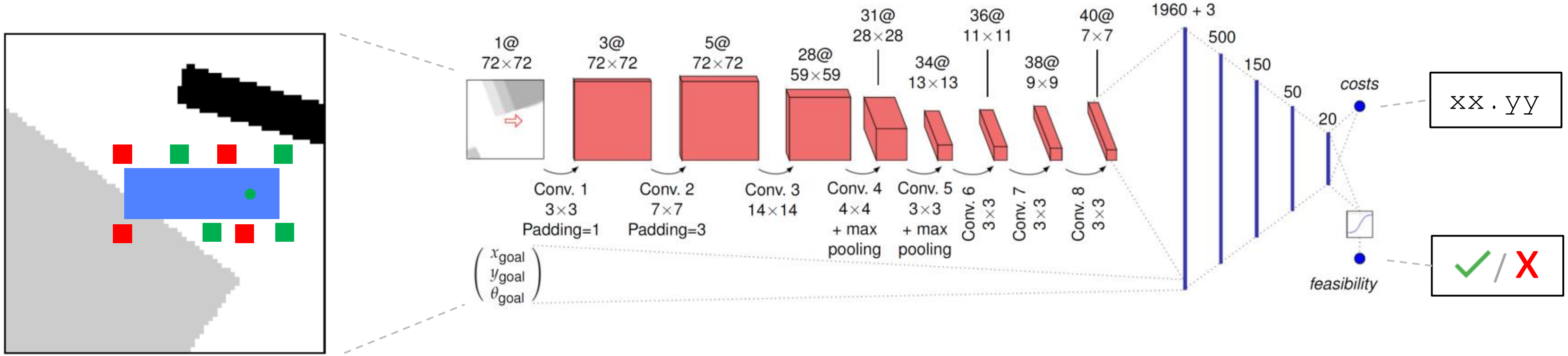
Planning problem



[Klamt and Behnke, ICRA 2019]

Abstraction CNN

- Predict feasibility and costs of local detailed planning

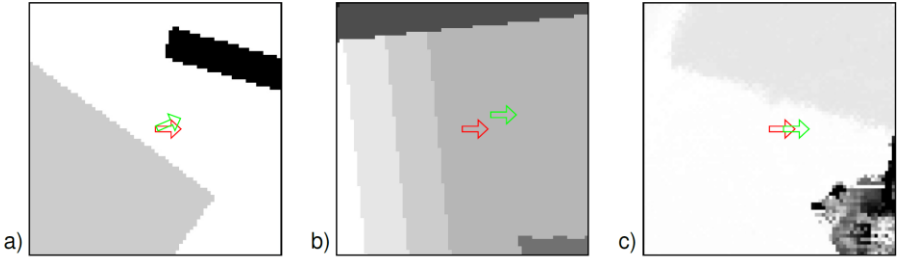


Training data
- generated with random obstacles, walls, staircases
- *costs* and *feasibility* from detailed A*-planner
- ~250.000 tasks

[Klamt and Behnke, ICRA 2019]

Learned Cost Function: Abstraction Quality

- CNN predicts feasibility and costs better than manually tuned geometric heuristics

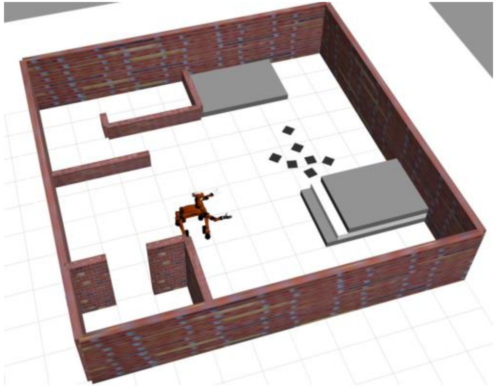


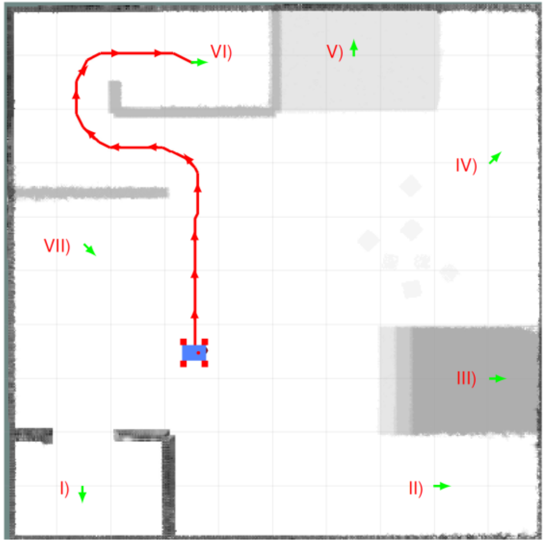
	random	*simulated*	*real*
feasibility correct, man.tuned	79.27%	65.35%	69.77%
Error($\mathcal{C}_{\mathrm{a,man.tuned}}$)	0.057	0.021	0.103
feasibility correct, CNN	95.04%	96.69%	92.62%
Error($\mathcal{C}_{\mathrm{a,CNN}}$)	0.027	0.013	0.081

[Klamt and Behnke, ICRA 2019]

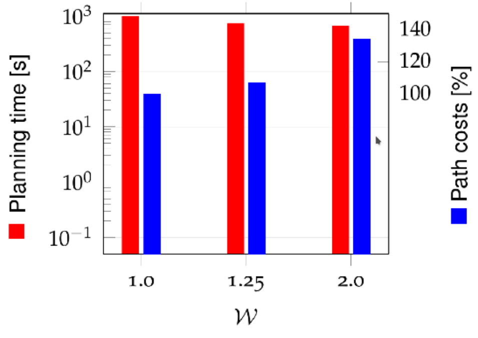
Experiments – Planning Performance

- Learned heuristics accelerates planning, without increasing path costs much

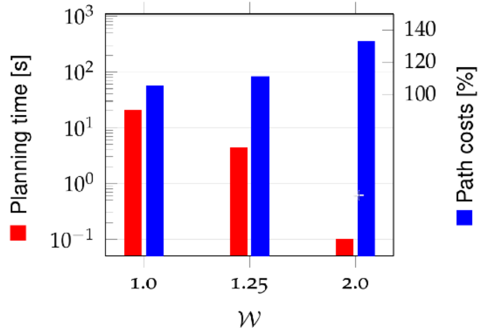




Geometric heuristic



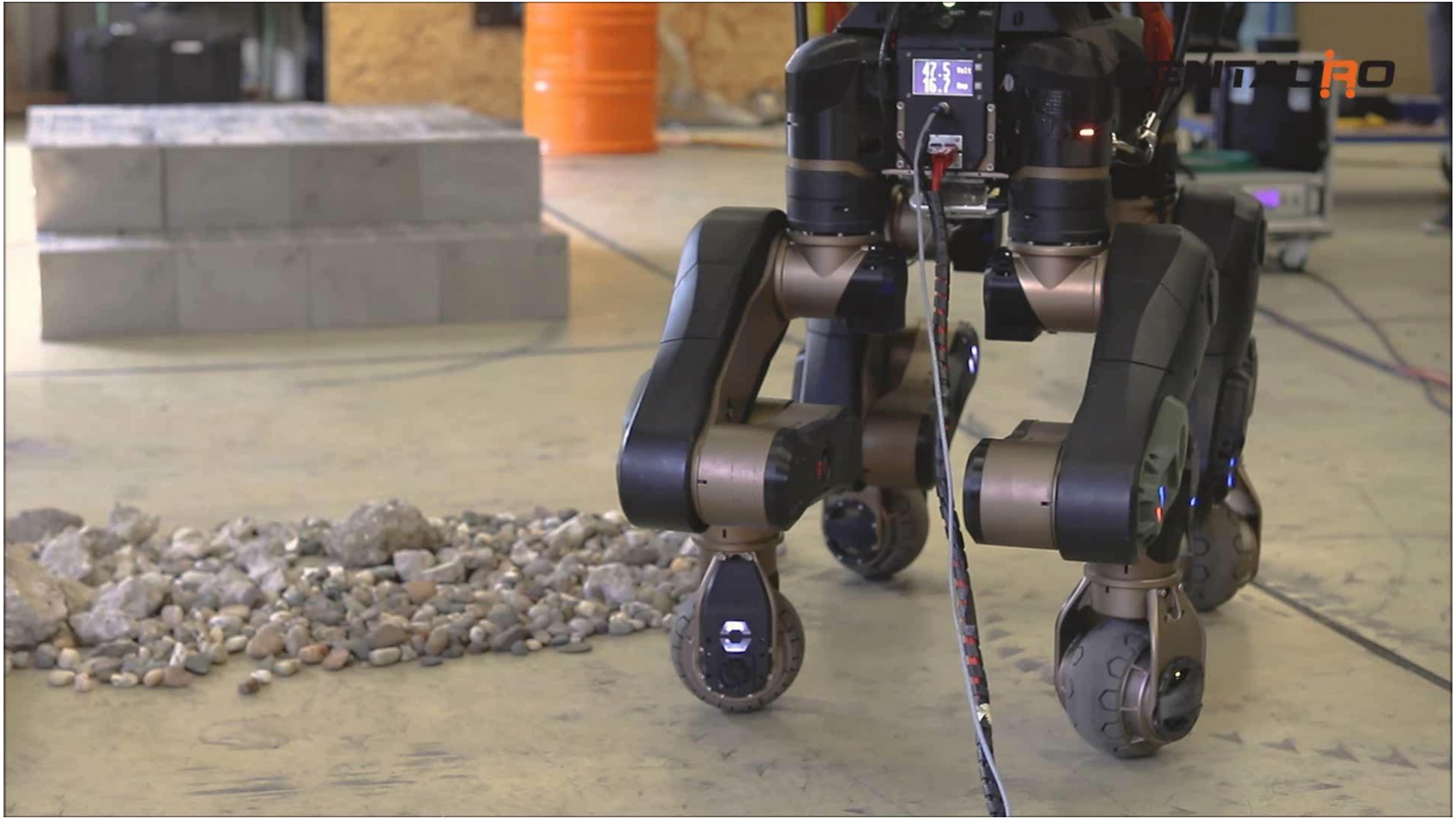
Abstract representation heuristic



Heuristic preprocessing: 239 sec

[Klamt and Behnke, ICRA 2019]

CENTAURO Evaluation @ KHG: Locomotion Tasks



[Klamt et al. RAM 2019]

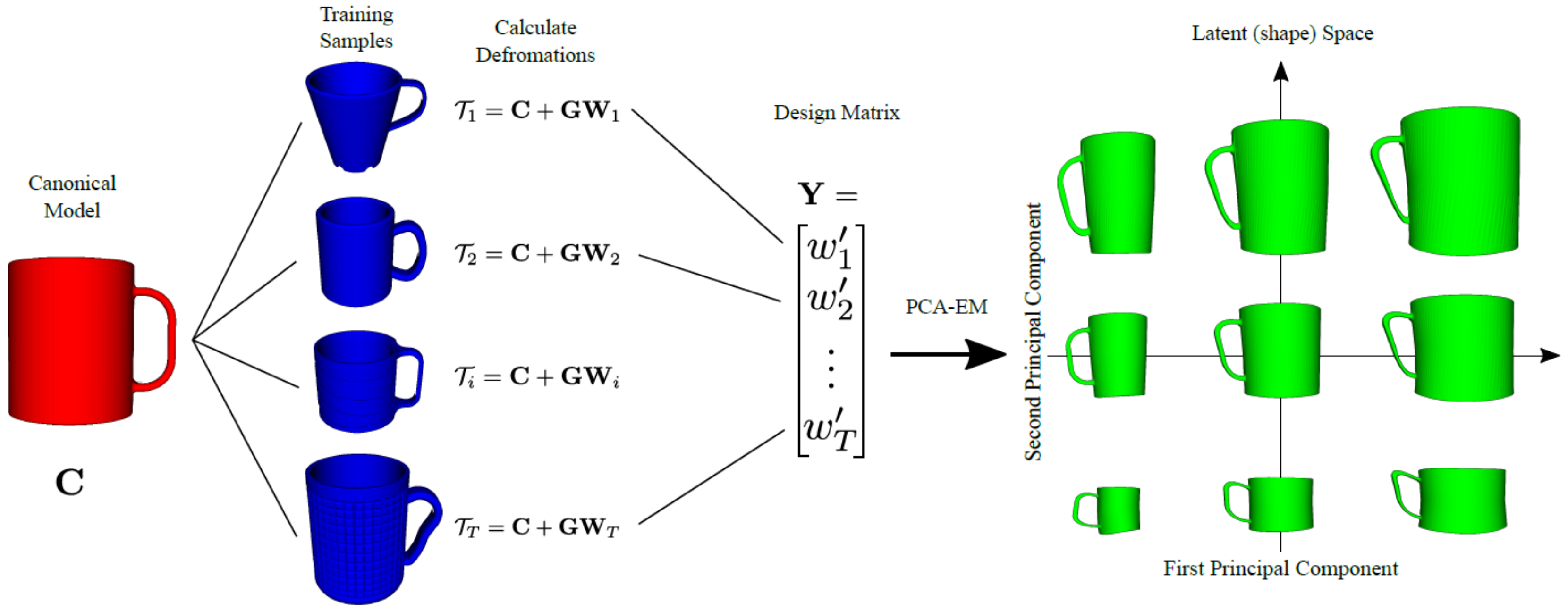
Transfer of Manipulation Skills

Knowledge Transfer

[Rodriguez and Behnke ICRA 2018]

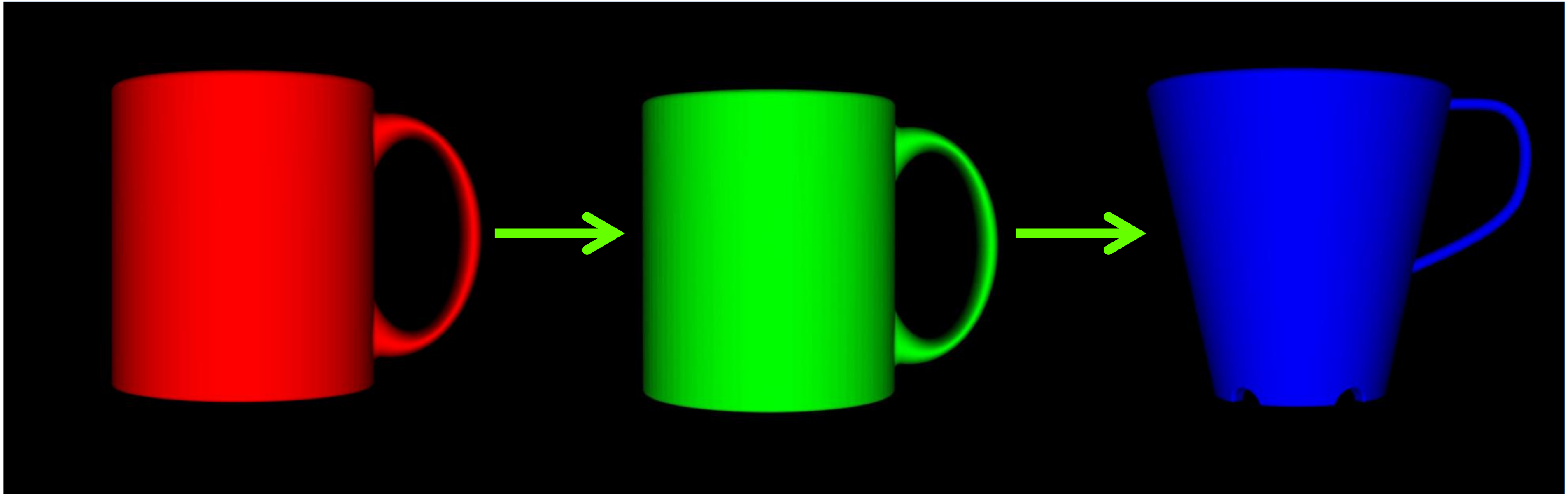
Learning a Latent Shape Space

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



[Rodriguez and Behnke ICRA 2018]

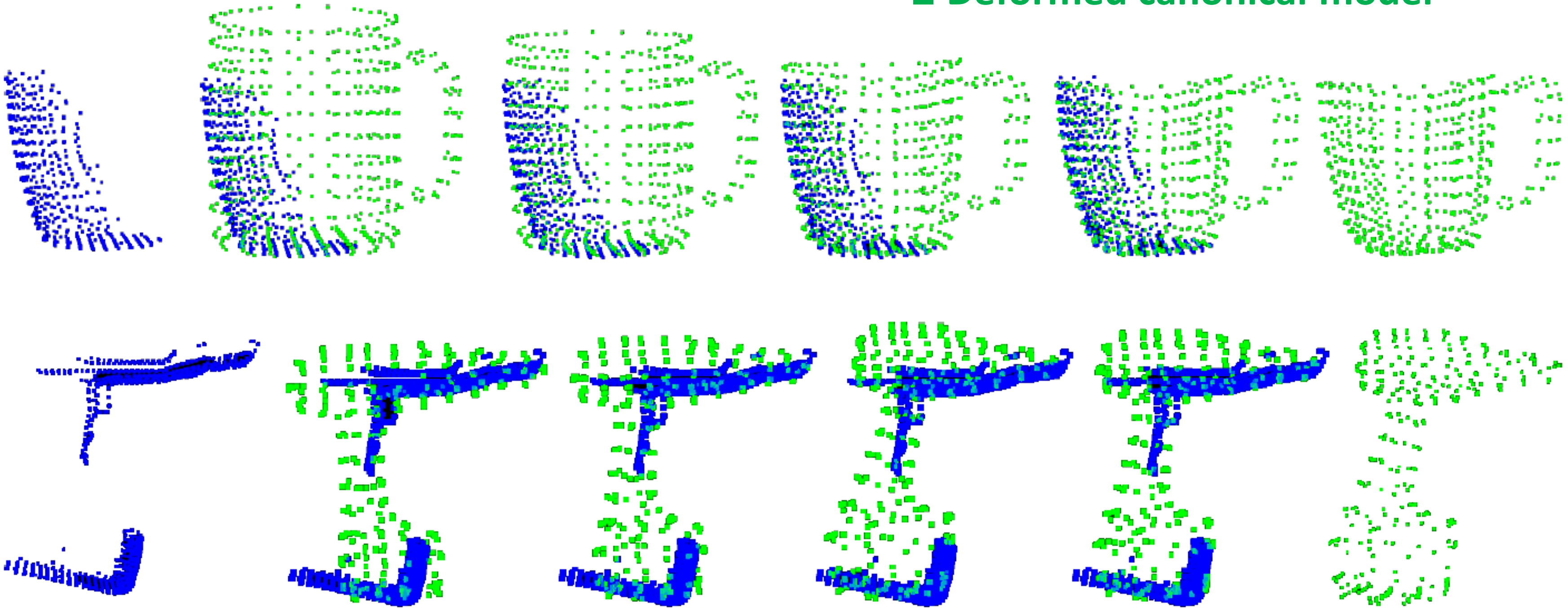
Interpolation in Shape Space



[Rodriguez and Behnke ICRA 2018]

Shape-aware Non-rigid Registration

■ Partial view of novel instance
■ Deformed canonical model

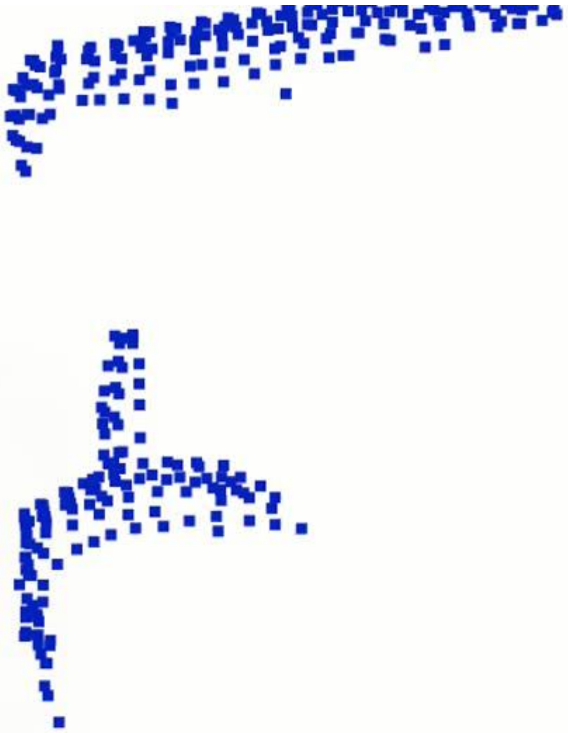


[Rodriguez and Behnke ICRA 2018]

Shape-aware Registration for Grasp Transfer

- Full point cloud

- Partial view

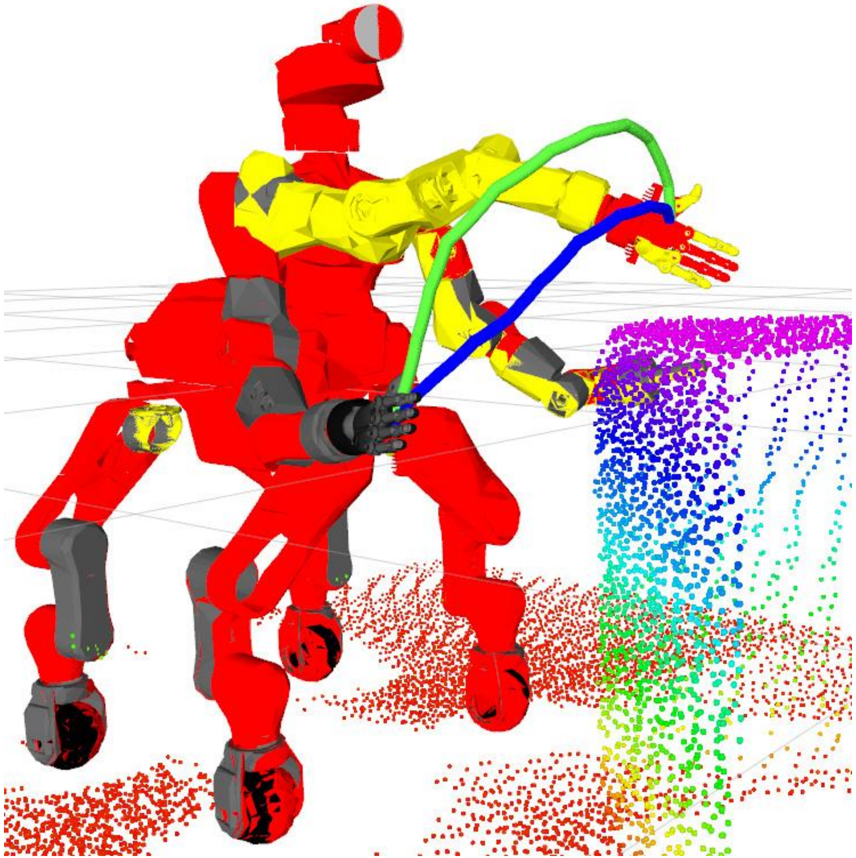


[Rodriguez and Behnke ICRA 2018]

Collision-aware Motion Generation

Constrained Trajectory Optimization:

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

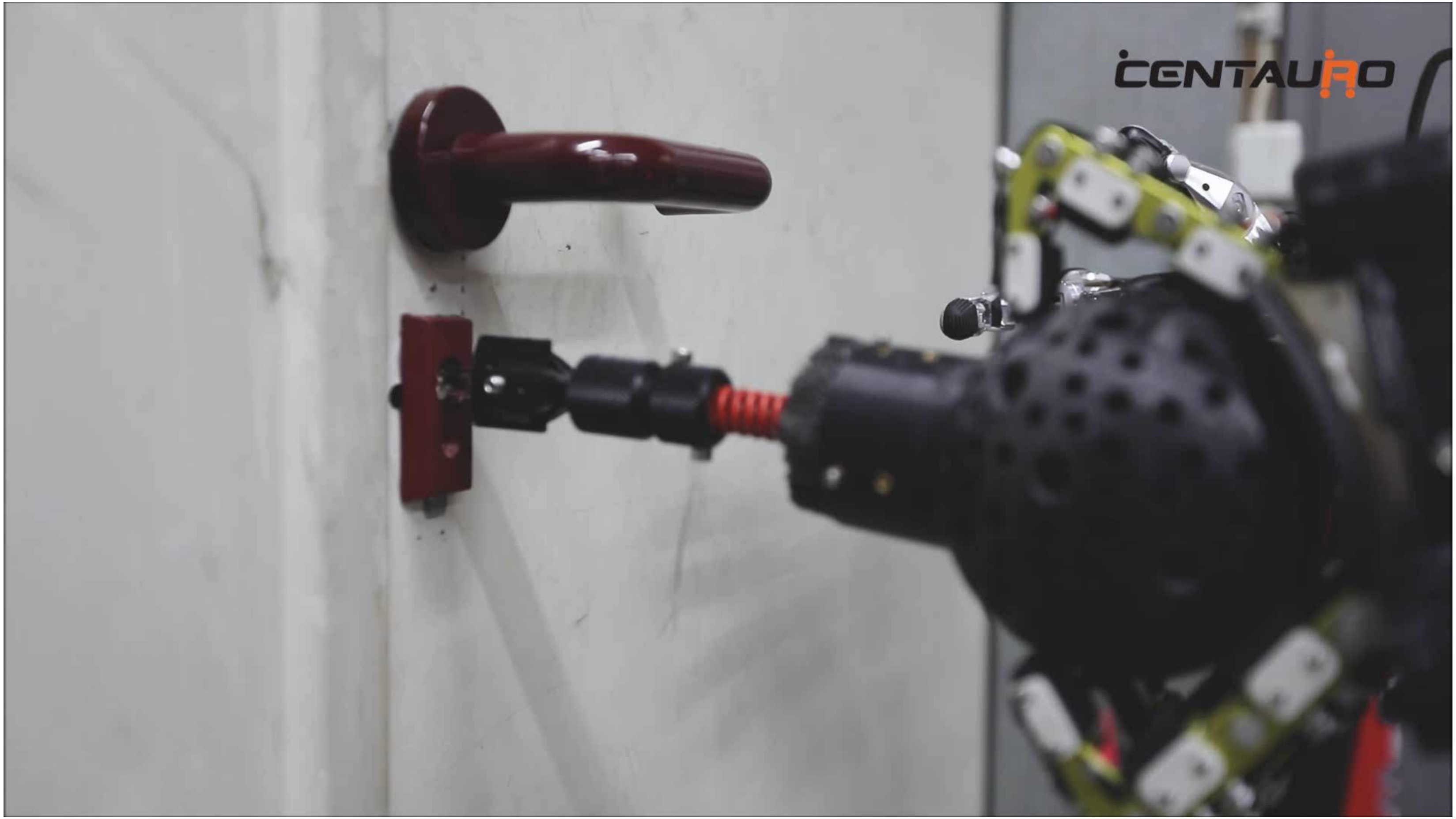


[Pavlichenko et al., IROS 2017]

Grasping an Unknown Power Drill and Fastening Screws

[Rodriguez and Behnke ICRA 2018]

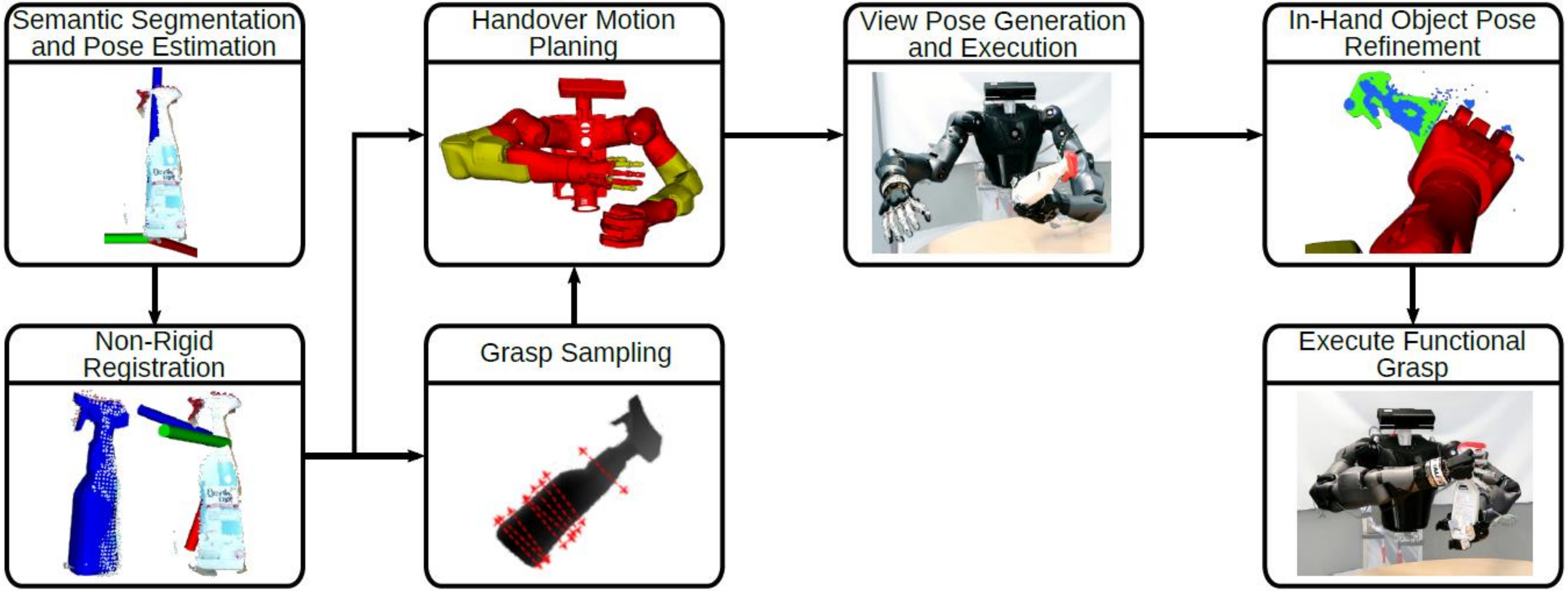
CENTAURO: Complex Manipulation Tasks



[Klamt et al. RAM 2019]

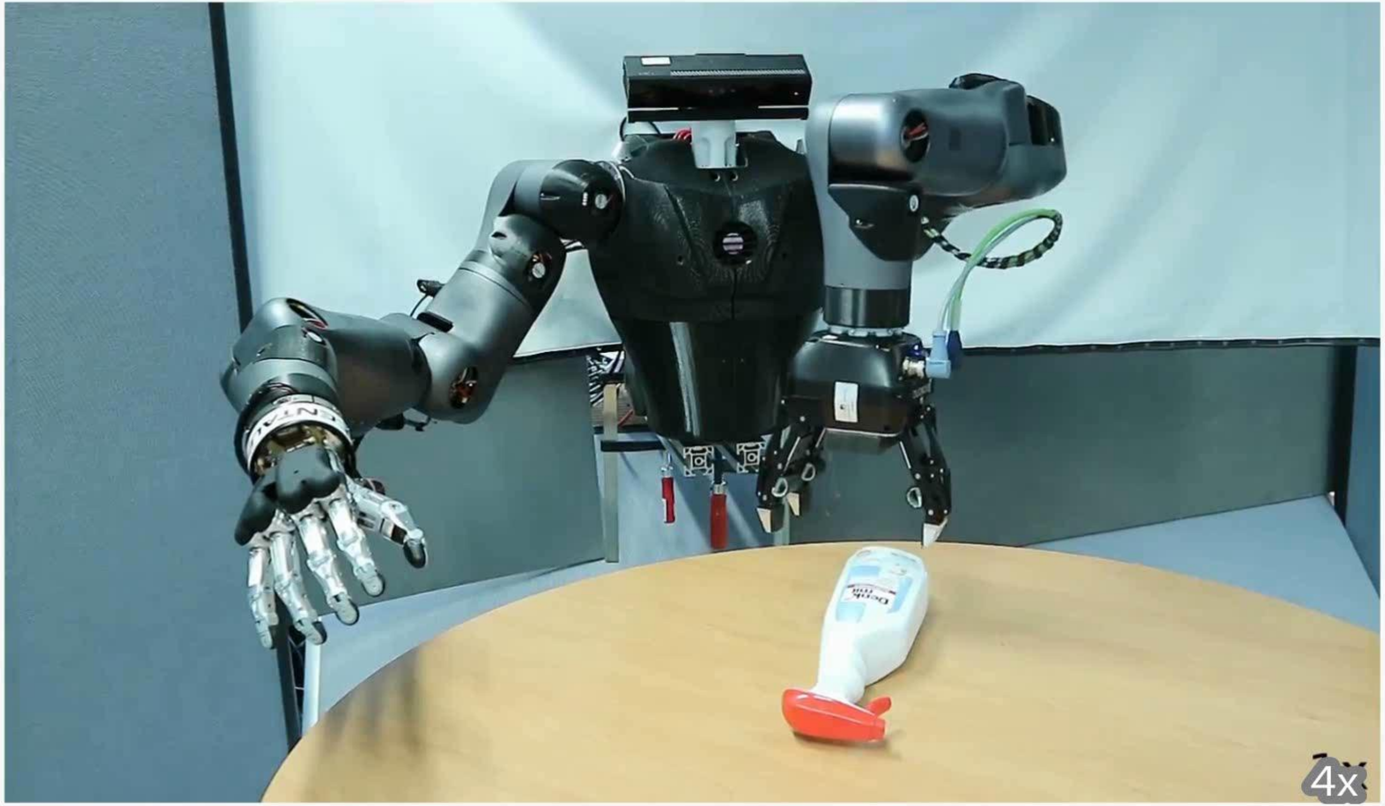
Regrasping for Functional Grasp

- 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]

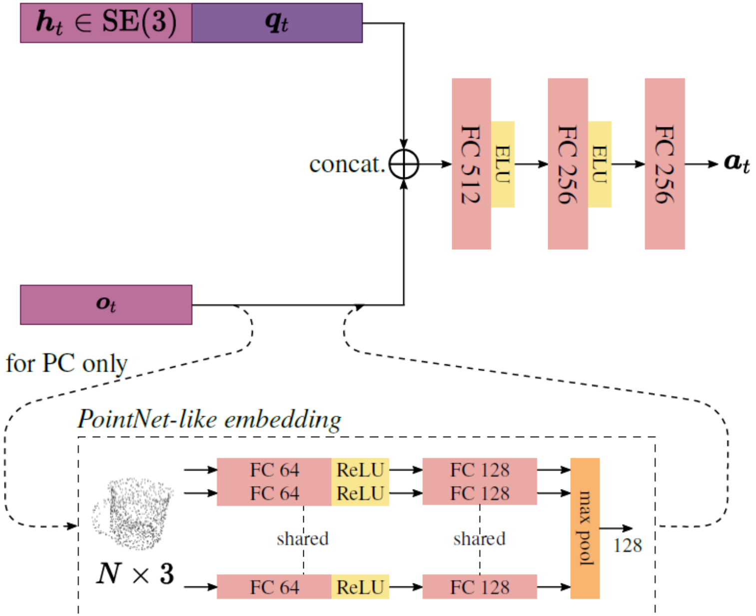
Regrasping Experiments

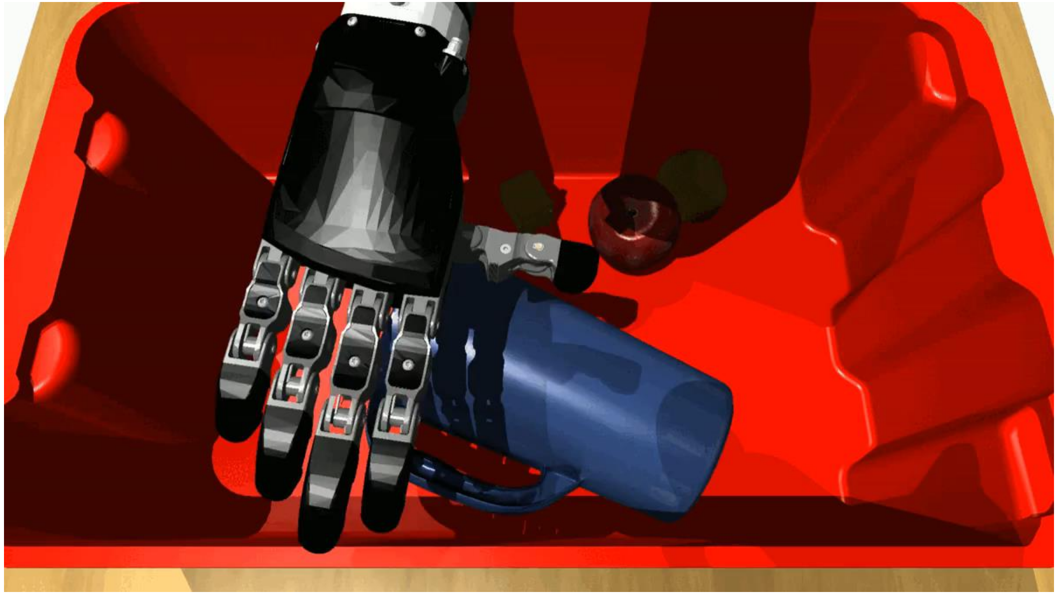


[Pavlichenko et al. Humanoids 2019]

Learning Interactive Grasping

- Deep RL-based interactive policy
- Input: object parameters or point cloud + hand pose
- Output: increments of hand DoF

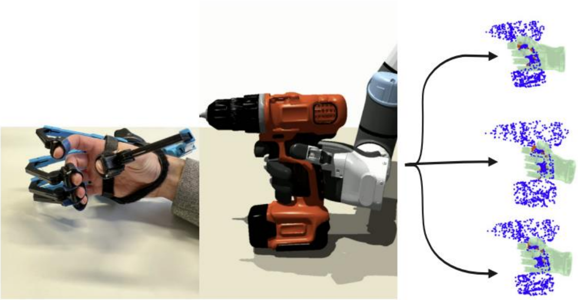




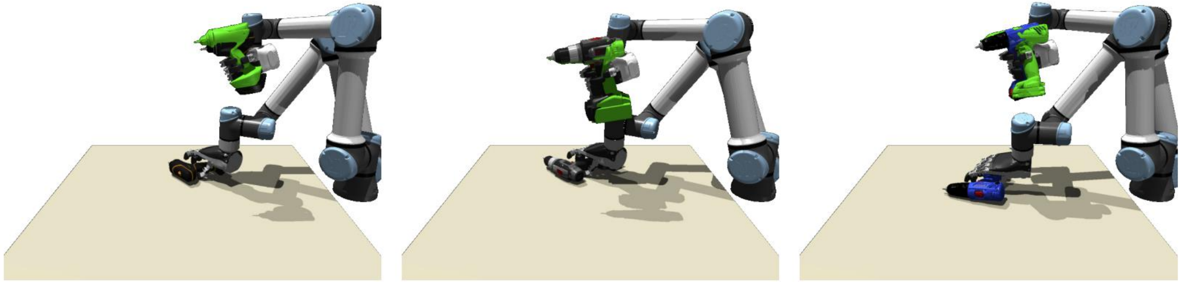
[Mosbach and Behnke, IRC 2022]

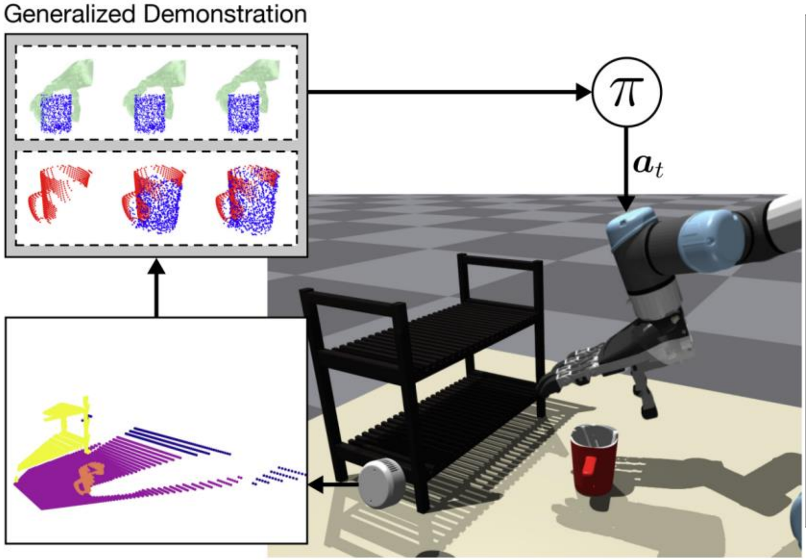
Learning Interactive Functional Grasping

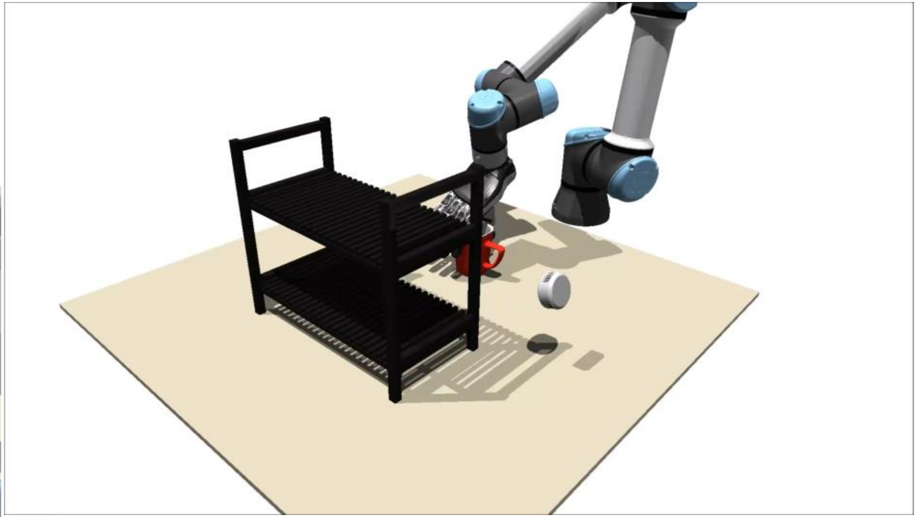
Generalization of a single demonstration



Interactive operation of unseen tools

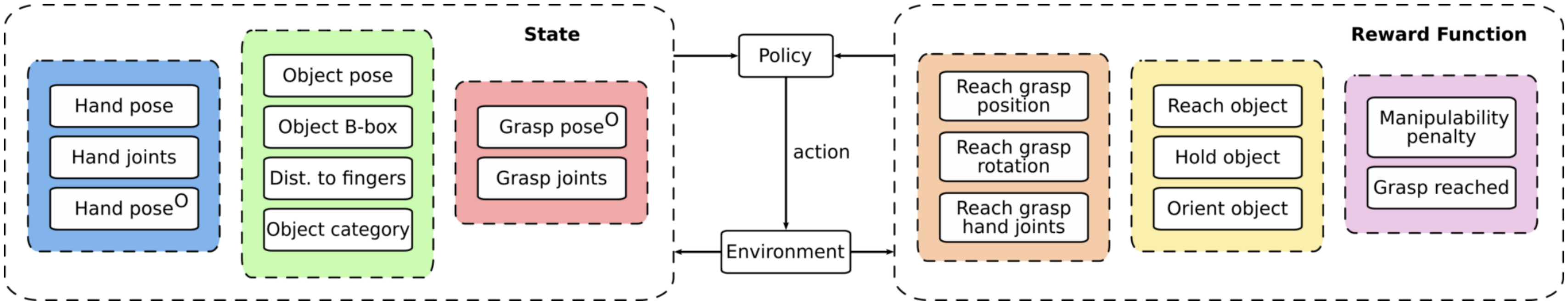




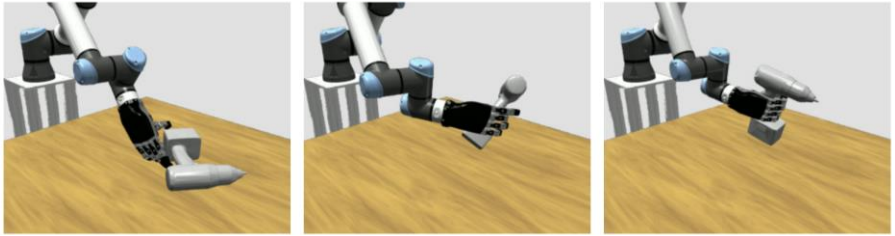


[Mosbach and Behnke CASE 2023, Best Paper Award]

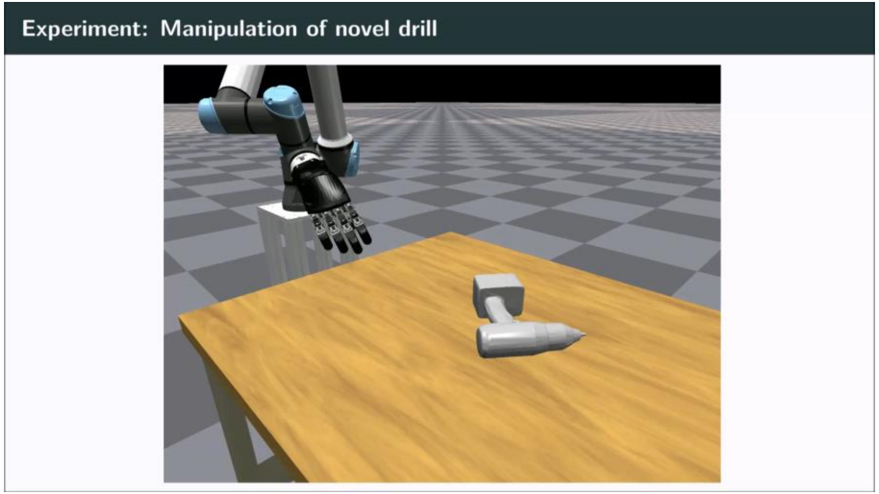
Learning Pre-grasp Manipulation for Human-like Functional Grasping



- Dense multi-component reward function encodes desired functional grasp



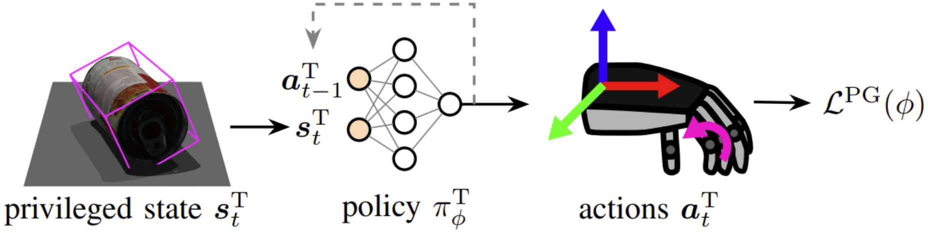
- Learns to reposition and reorient objects to achieve functional grasps



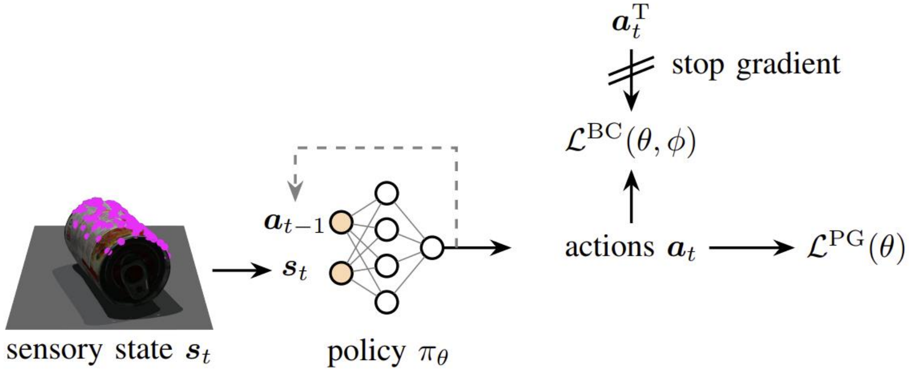
[Pavlichenko and Behnke CASE 2023]

Grasp Anything: Augmenting Reinforcement Learning with Instance Segmentation to Grasp Arbitrary Objects

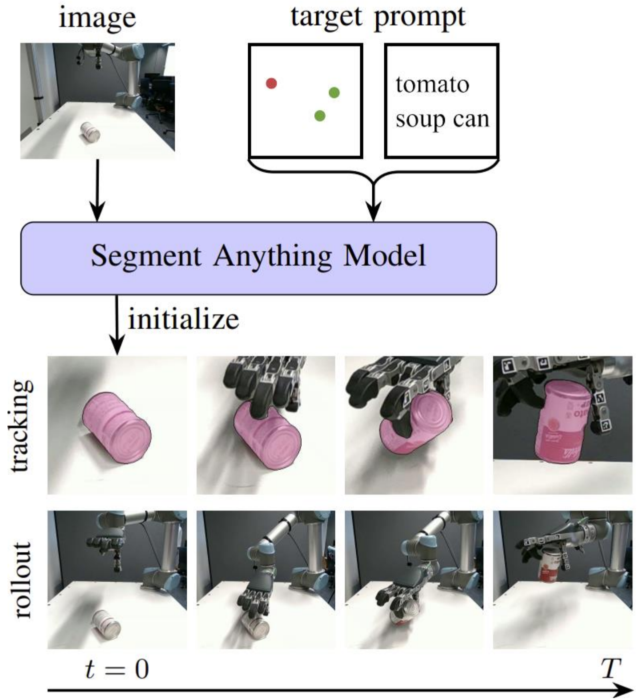
- Teacher training



- Teacher-guided sensorimotor learning



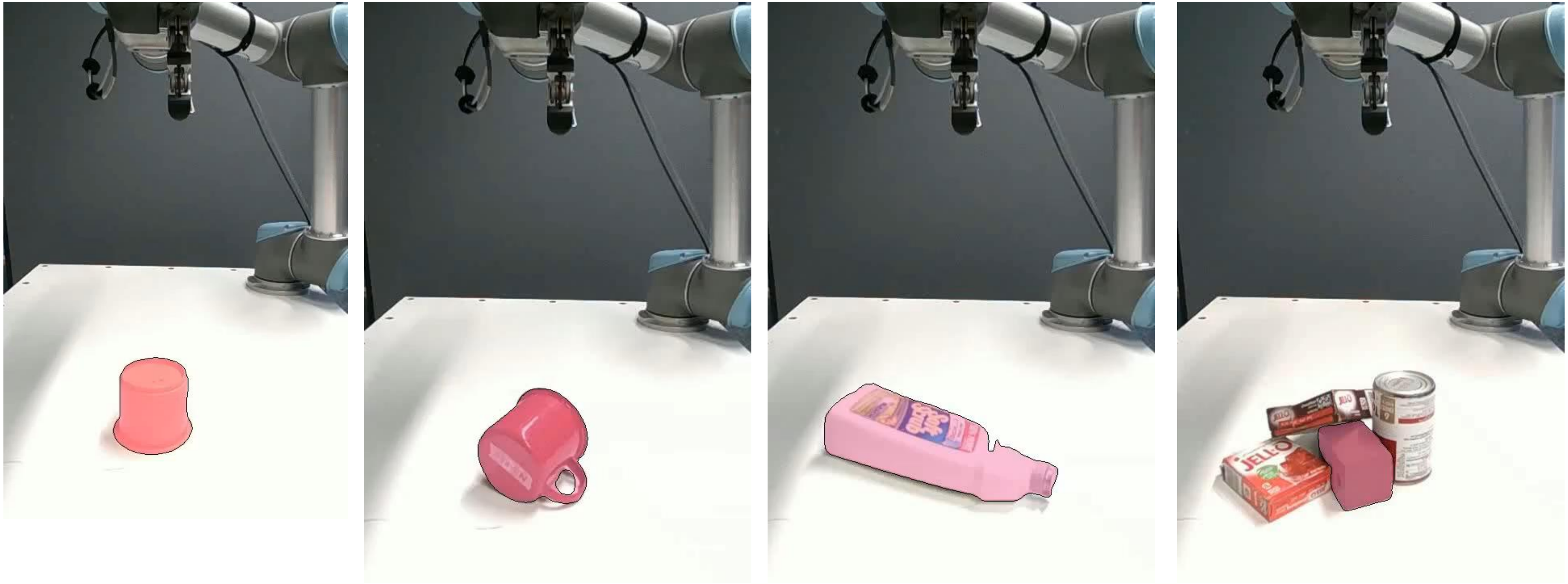
- Real-world deployment of promptable grasping policy



[Mosbach and Behnke ICRA 2024]

Grasp Anything: Augmenting Reinforcement Learning with Instance Segmentation to Grasp Arbitrary Objects

- Learned policy with improved object visibility is real-world deployable



[Mosbach and Behnke ICRA 2024]

NimbRo@Home RoboCup 2023 in Bordeaux, France

LLM Integration

- Utilization of Function Calling Models (GPT / Mistral)
- Map available robot functionalities to user commands
- Given Information:
  - Environment info
  - Callable robot functionalities
  - Can request and query additional image data (counting, descriptions, ...)

User: Please bring me a sponge from the kitchen.
System: Briefly summarize the next steps to accomplish the task in text form.
Assistant:
1. Drive to the kitchen.
2. Search for a sponge in the kitchen.
3. Grasp the sponge.
4. Drive back to the parlor.
5. Hand over the sponge to the user.
6. Call the exit function to complete the interaction.

System: From now on only use function calls to execute the task.
Assistant: drive_to_location: {"location": "kitchen"}
Tool: You successfully arrived in the new location kitchen.
Assistant: find_object: {"object_name_list": ["sponge"]}
Tool: The following items were found in the kitchen: 3 Sponges
Assistant: grasp_object: {"object_name": "sponge"}
Tool: You successfully grasped the object sponge.
Assistant: drive_to_location: {"location": "parlor"}
Tool: You successfully arrived in the new location parlor.
Assistant: place_object: {"object_name": "sponge"}
Tool: You successfully placed the object sponge.
Assistant: exit: {}

Cooking Advice

- Q: Please describe the content of this image. Focus your description on food items, drinks, and ingredients that could be used for cooking.

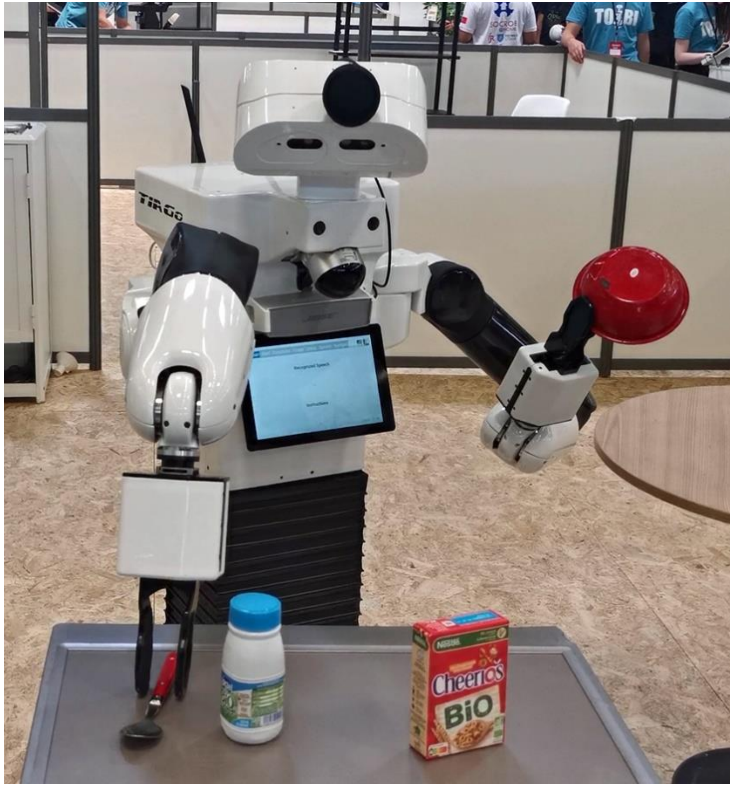
- A: The image shows a variety of food items and drinks arranged on a white surface. There is a can of SPAM, which is a type of canned cooked pork product. Next to it is an open can of Campbell's Tomato Soup...

Cooking Advice

- Q: How can I cook dinner from these ingredients? <PROMPT>
- A: You can create a simple and creative meal using these ingredients. Here's a suggestion:
  - SPAM and Orange Glaze: Slice the SPAM into thin pieces and fry them in a pan until they are crispy on both sides. Create a glaze by squeezing the juice of the orange into a small saucepan, adding a bit of the Coca-Cola Zero Sugar for sweetness, and reducing it over medium heat until it thickens. Pour the glaze over the fried SPAM slices.
  - Tomato Soup Enhancement: Warm the Campbell's Tomato Soup on the stove. To add more depth to the flavor, you could mix in a small amount of the orange juice and a pinch of the JELL-O powder for a hint of sweetness and tanginess.

Conclusions

- Developed capable robotic systems for challenging scenarios
  - Bin picking
  - Disaster response
  - Domestic service tasks
- Challenges include
  - 4D semantic perception
  - High-dimensional motion planning
  - Human-robot interaction
- Promising approaches
  - Prior knowledge (pretrained models, inductive bias, LLMs)
  - Shared experience (fleet learning)
  - Shared autonomy (human-robot)
  - Instrumented environments



UNIVERSITÄT BONN AIS