Imitating Task-oriented Grasps from Human Demonstrations with a Low-DoF Gripper

Timothy Patten and Markus Vincze

Vision for Robotics
Automation and Control Institute
TU Wien, Vienna, Austria

Contact: patten@acin.tuwien.ac.at

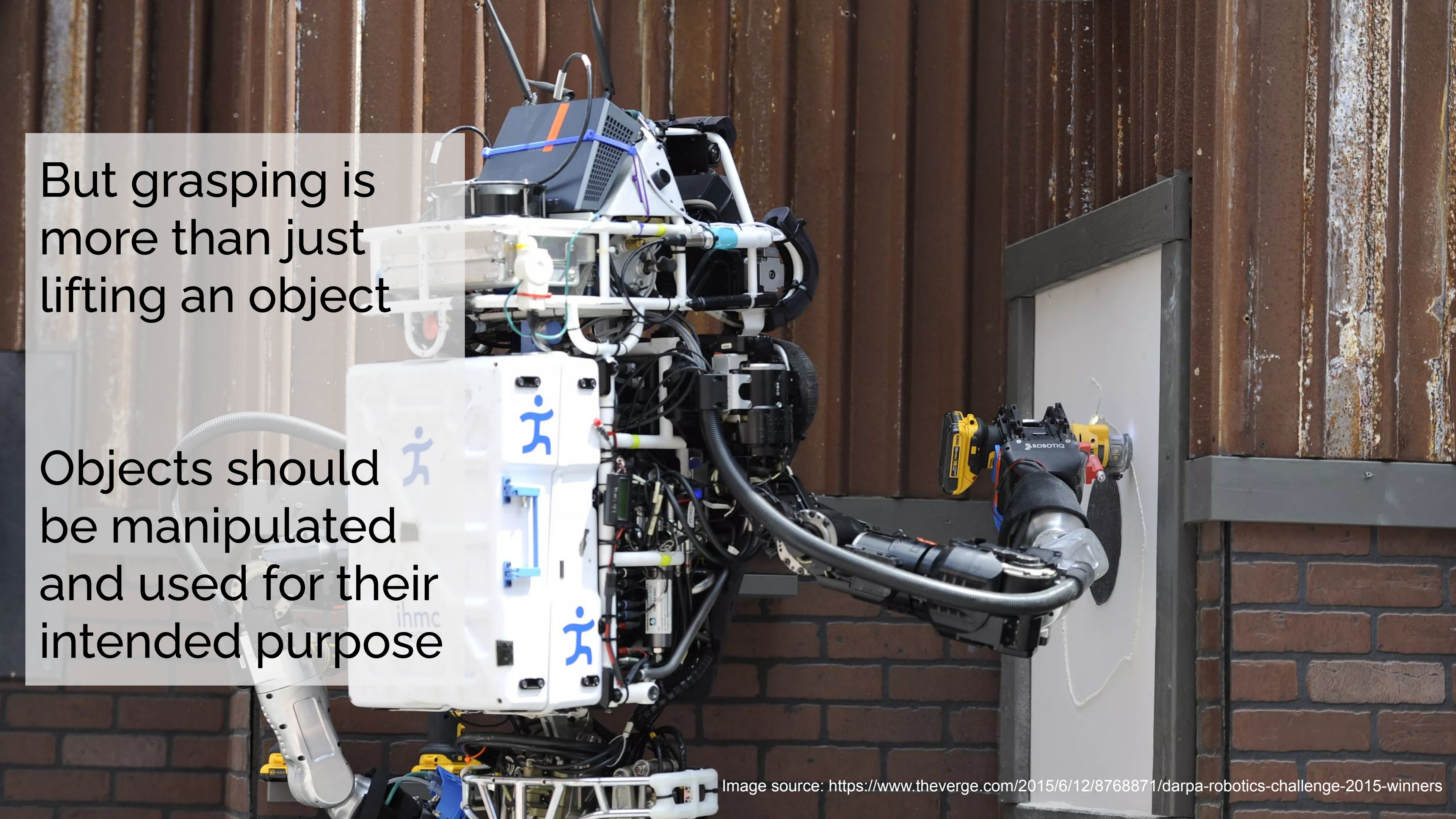
The Sixteenth International Conference on Autonomic and Autonomous Systems
ICAS 2020
September 27, 2020 to October 01, 2020 - Lisbon, Portugal

Dr. Timothy Patten received his PhD from the Australian Centre for Field Robotics at the University of Sydney, Australia. He is now a postdoctoral researcher with the Vision for Robotics laboratory at the Technical University of Vienna (TU Wien), Austria.

He has been involved in a number of research and industry sponsored projects in which he worked on object segmentation, recognition, grasping and task planning. Currently, he is the principle investigator at TU Wien in the CHIST-ERA project InDex, for which he is developing methods for object tracking and semantic grasping.

Robotic grasping typically focuses on achieving a stable grasp for object transportation (e.g., bin picking)

Image source: https://www.aidanscannell.com/project/amazon-picking-challenge/

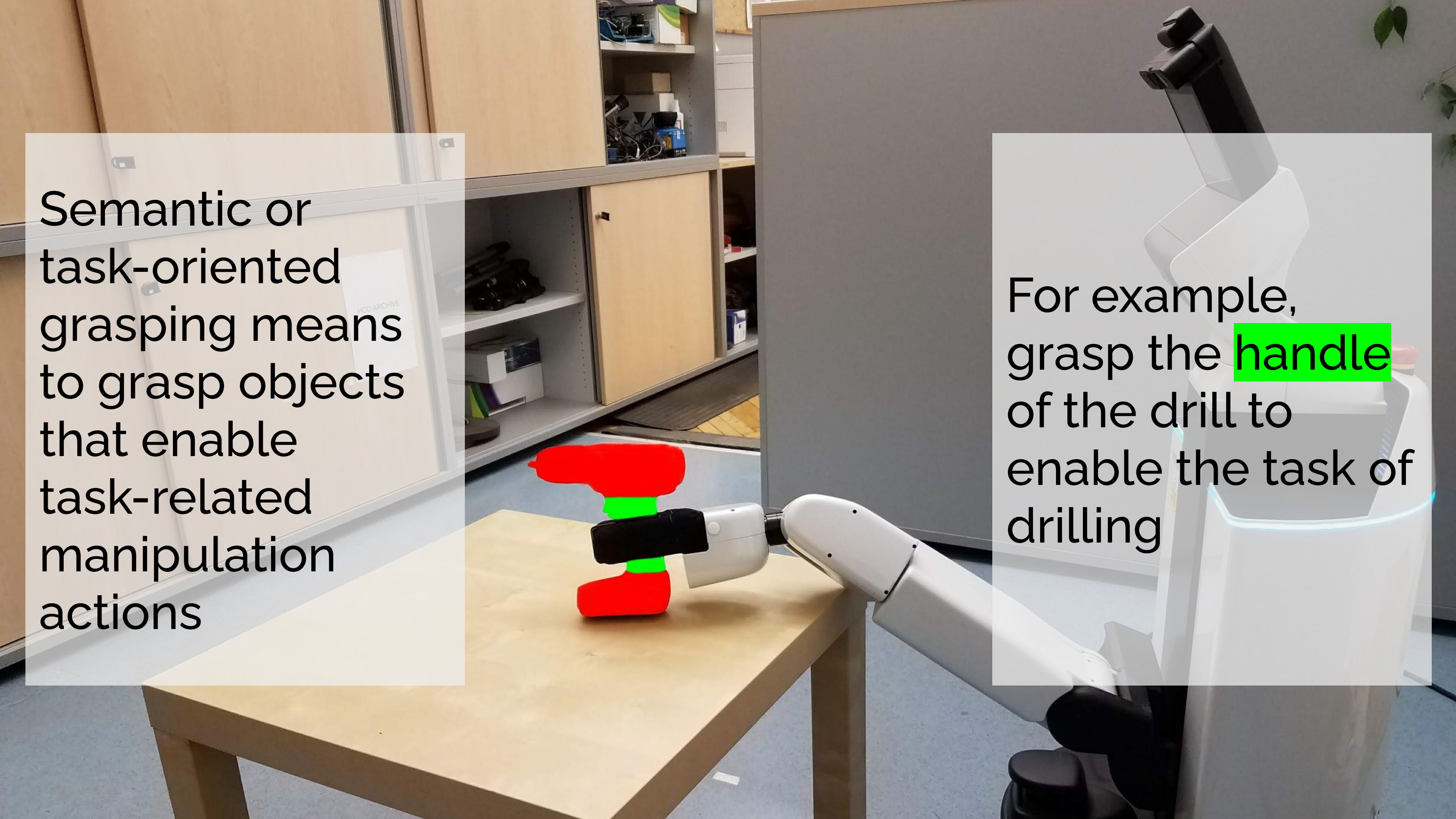


But grasping is more than just lifting an object

Objects should be manipulated and used for their intended purpose

Image source: https://www.theverge.com/2015/6/12/8768871/darpa-robotics-challenge-2015-winners

Semantic or task-oriented grasping means to grasp objects that enable task-related manipulation actions



Semantic or task-oriented grasping means to grasp objects that enable task-related manipulation actions

For example, grasp the handle of the drill to enable the task of drilling

Humans hold the key for semantic knowledge, therefore, learning from humans is a practical solution

M. Hirschmanner, C. Tsiourti, T. Patten and M. Vincze, “Virtual reality teleoperation of a humanoid using markerless human upper body pose imitation,” IEEE-RAS Humanoids, 2019, pp. 259–2650

Our Contributions

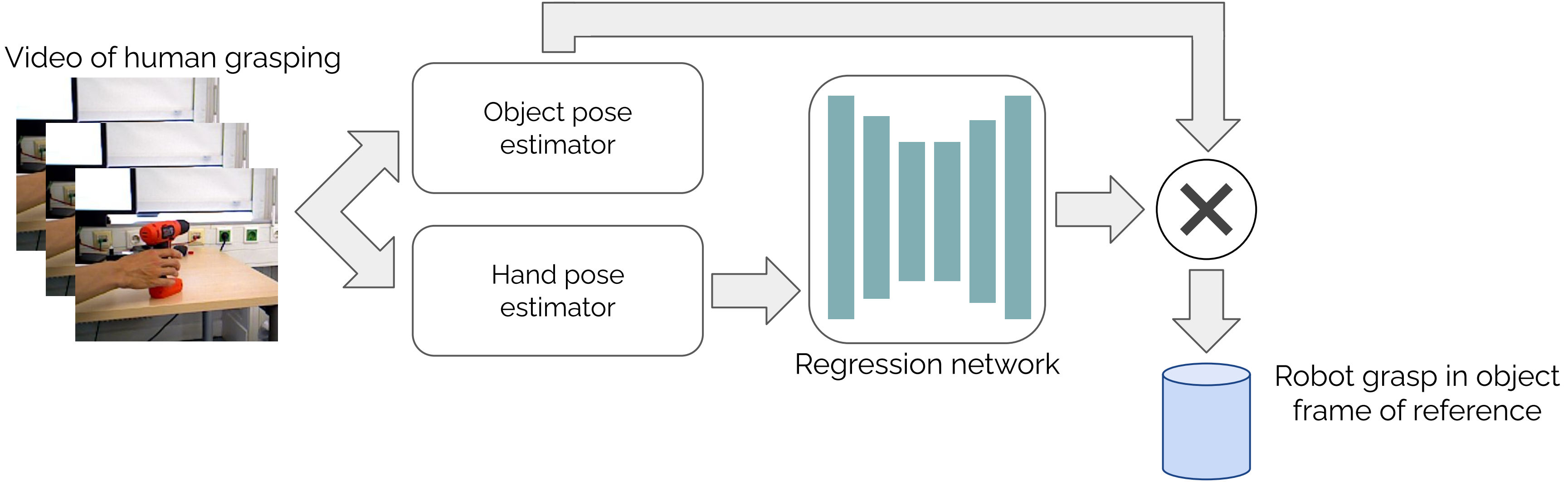
- A framework for imitating task-oriented grasps demonstrated by humans
    - Vision-based: requires no special instrumentation [1], manual annotation [2], physical interaction [3] or offline learning process [4]

- A neural network architecture to regress grasp parameterisation of a low-DoF (parallel-jaw) robotic gripper from human hand configuration

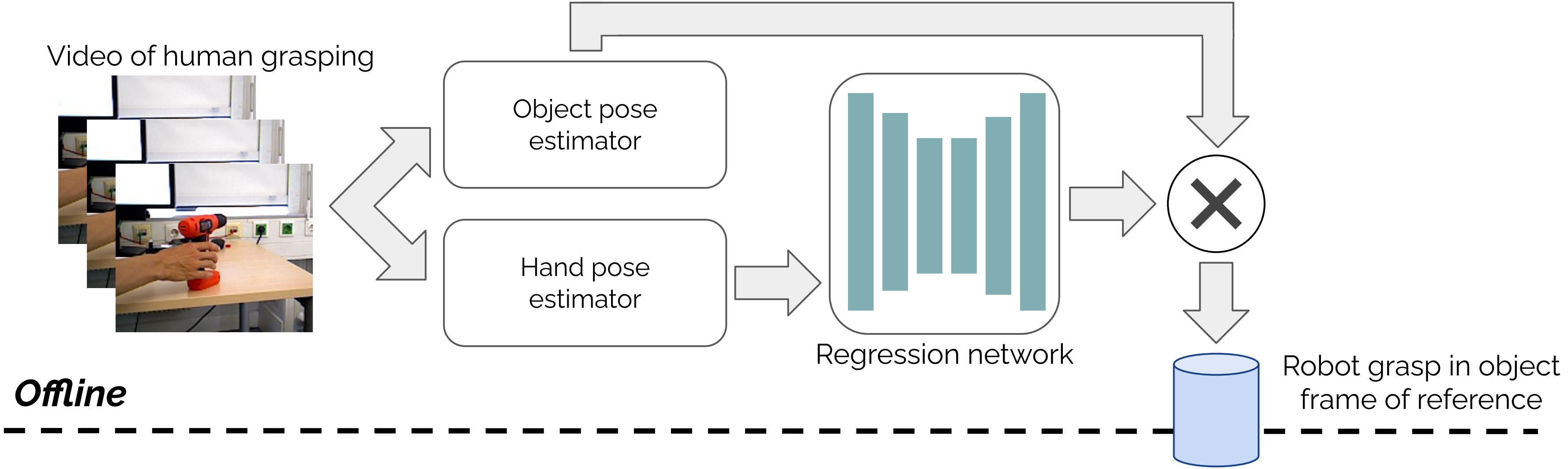
- Evaluation of grasp regression network and experiments of real-world task-oriented grasping with mobile manipulator

[1] J. Aleotti and S. Caselli, “Part-based robot grasp planning from human demonstration,” IEEE ICRA, 2011, pp. 4554–4560
[2] M. Hjelm, C. H. Ek, R. Detry, and D. Kragic, “Learning human priors for task-constrained grasping,” ICVS, 2015, pp. 207–217
[3] S. H. Kasaei, N. Shafii, L. S. Lopes, and A. M. Tomé, “Interactive open-ended object, affordance and grasp learning for robotic manipulation,” IEEE ICRA, 2019, pp. 3747–3753
[4] D. Antotsiou, G. Garcia-Hernando, and T.-K. Kim, “Task-oriented hand motion retargeting for dexterous manipulation imitation,” ECCV Workshops, 2018

Grasp Imitation Framework



Grasp Imitation Framework



Video of human grasping

Object pose estimator

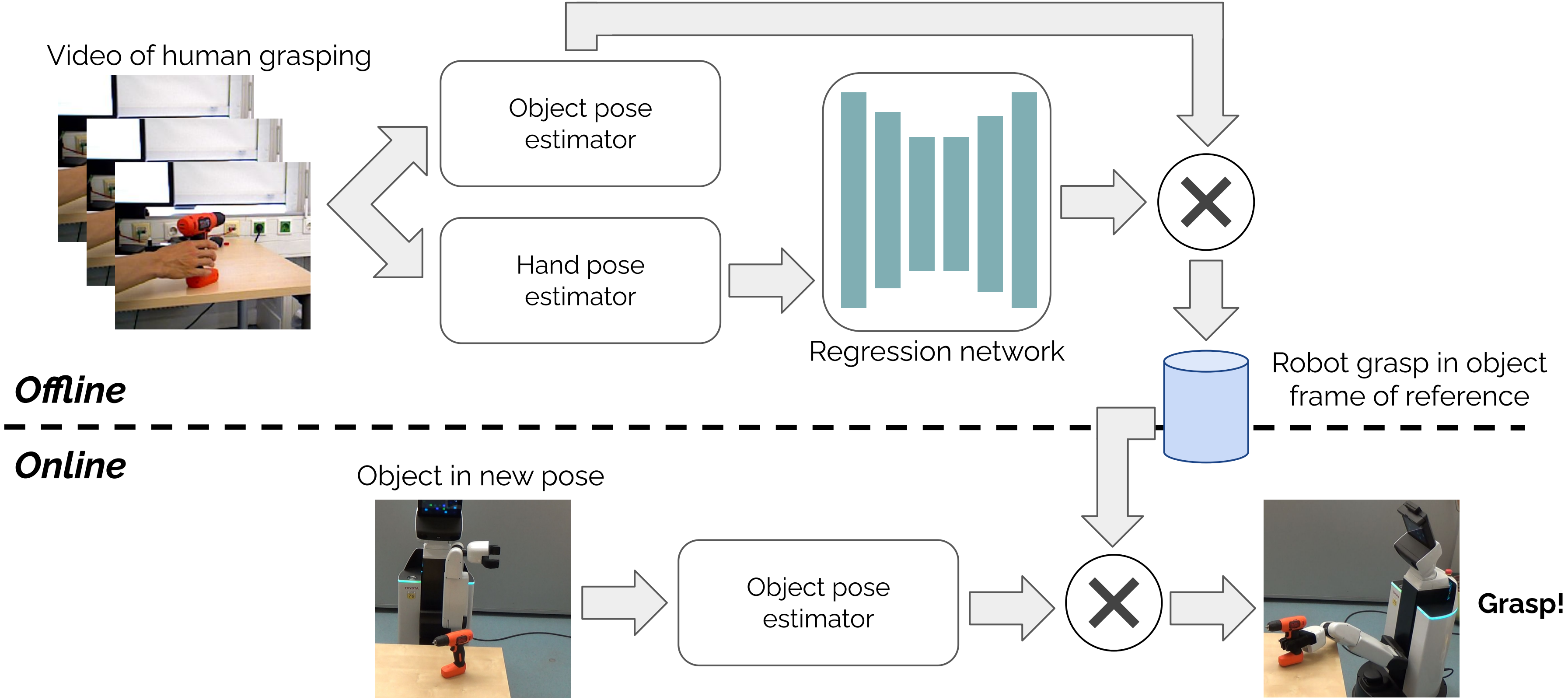
Hand pose estimator

Regression network

Robot grasp in object frame of reference

Offline

Grasp Imitation Framework



Video of human grasping

Object pose estimator

Hand pose estimator

Regression network

Robot grasp in object frame of reference

Offline

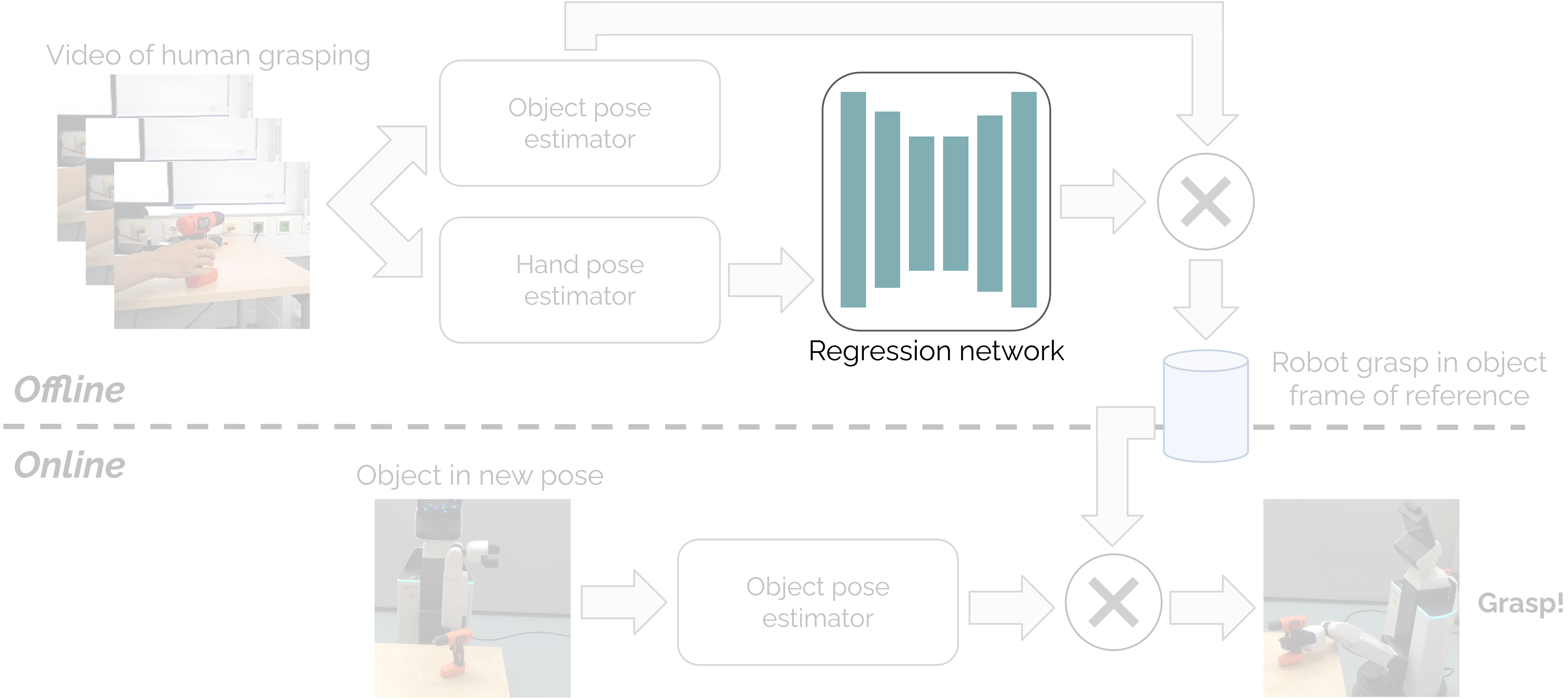
Online

Object in new pose

Object pose estimator

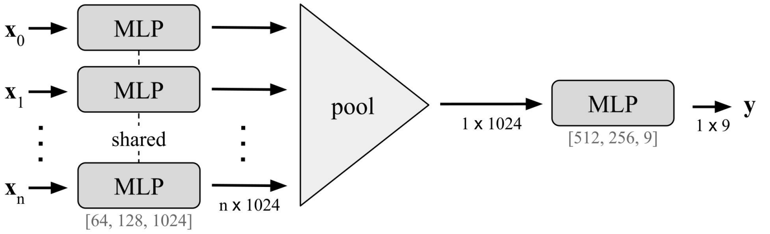
Grasp!

Grasp Imitation Framework

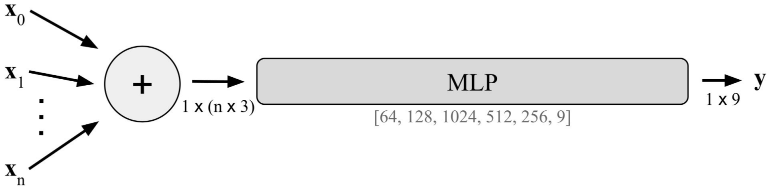


Robotic Grasps from Human Hand Configurations

- Mapping from human hand $\mathbf{H} \in \mathbb{R}^H$ to robot grasp $\mathbf{G} \in \mathbb{R}^G$, i.e. $\mathbf{G} = \mathcal{F}(\mathbf{H})$
  - We model function $\mathcal{F}$ as a neural network
- Baseline architecture inspired by PointNet [1]
  - Each point on hand input to a multilayer perception (MLP)
  - Feature maps transformed to global feature with pooling operation
  - Global feature passes through another MLP
  - Output the translation (x, y, z), approach angle (a_x, a_y, a_z) and closing angle (c_x, c_y, c_z)



- Architecture exploiting known joint configuration
  - Sorted point coordinates concatenated and passes through a single MLP
  - Same output as baseline

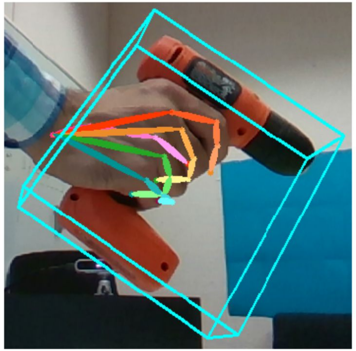


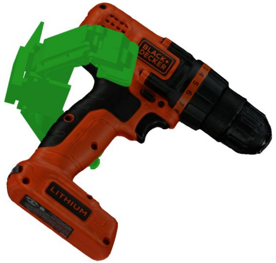
[1] C. R. Qi, H. Su, M. Kaichun, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” IEEE/CVF CVPR, 2017, pp. 77–85

Network Training

- Loss function
  - *l2* loss for the normalised values of the translation and angles: $\mathcal{L}(\mathbf{y}^{GT}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i^{GT} - \mathbf{y}_i)^2$
  - Account for symmetry of parallel-jaw gripper; take minimum of prediction and flipped version (rotated 180° around closing angle): $\mathcal{L}_{\text{sym}}(\mathbf{y}^{\text{GT}}, \mathbf{y}) = \min(\mathcal{L}(\mathbf{y}^{\text{GT}}, \mathbf{y}), \mathcal{L}(\mathbf{y}^{\text{GT}}, \mathbf{y}^{\text{flipped}}))$

- Data from the HO3D dataset [1]
  - Sequences of people manipulating object in right hand
  - Objects from the YCB dataset [2]
  - Object pose and hand joints accurately annotated
  - Corresponding grasps were hand annotated for this work
  - Local and global augmentation applied

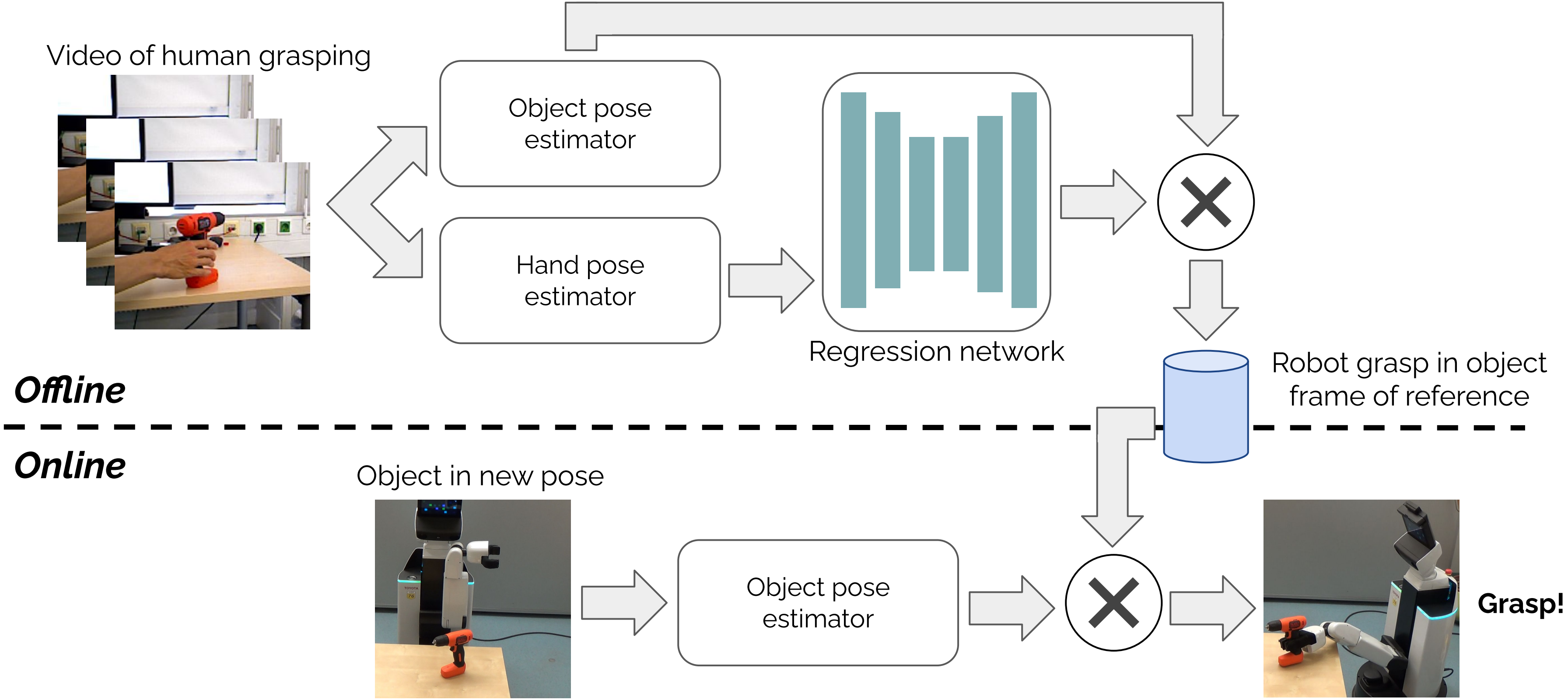




[1] S. Hampali, M. Oberweger, M. Rad, and V. Lepetit, “HOnnotate: A method for 3D annotation of hand and object poses,” IEEE/CVF CVPR, 2020, pp. 3196–3206

[2] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, “Yale-CMU-Berkeley dataset for robotic manipulation research,” IJRR, 36(3), 2017, pp. 261–268

Grasp Imitation Framework



Video of human grasping

Object pose estimator

Hand pose estimator

Regression network

Robot grasp in object frame of reference

Offline

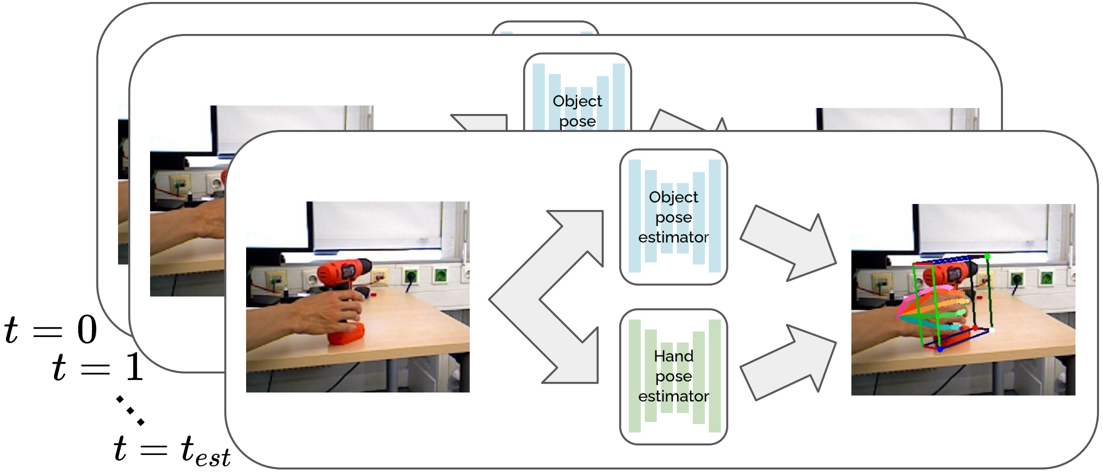
Online

Object in new pose

Object pose estimator

Grasp!

Offline Robot Grasp Pose Estimation



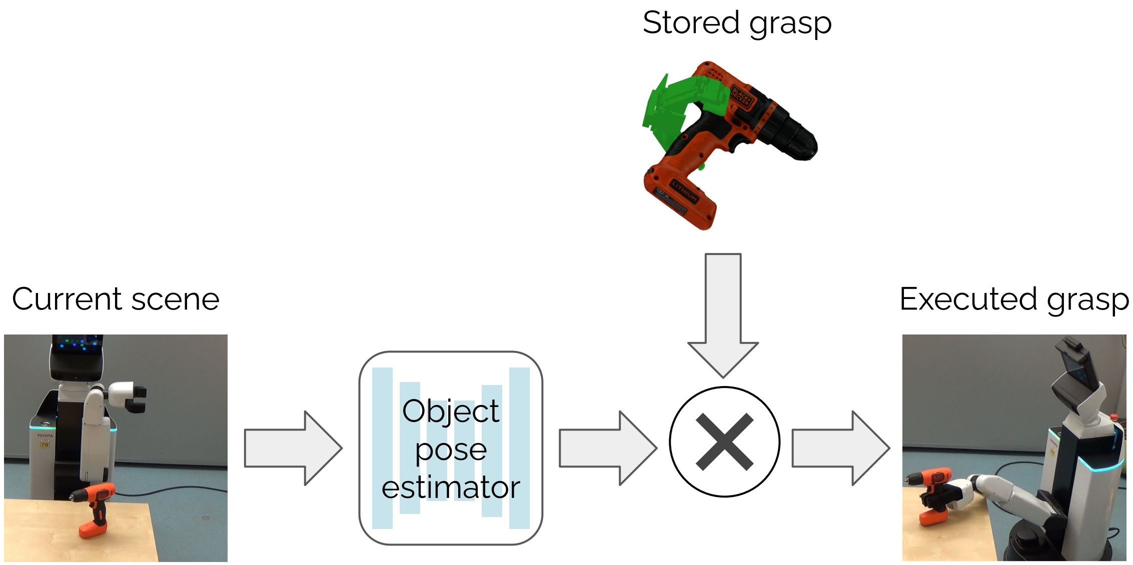
Object pose [1] estimated in each frame until it is observed to move (i.e., grasped and lifted)

This frame ($t = t_{est}$) used to estimate the hand pose [2] w.r.t. the object pose

Hand pose used to estimate the robot gripper pose with regression network

[1] K. Park, T. Patten, and M. Vincze, "Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation," IEEE ICCV, 2019, pp. 7668–7677
[2] P. Panteleris, I. Oikonomidis, and A. Argyros, "Using a single RGB frame for real time 3D hand pose estimation in the wild," IEEE WACV, 2018, pp. 436–445

Online Robot Grasp Execution



Stored grasp

Current scene

Object pose estimator

Executed grasp

Object pose [1] estimated in current scene

Then, transform demonstrated grasp to new scene and execute

[1] K. Park, T. Patten, and M. Vincze, "Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation," IEEE ICCV, 2019, pp. 7668–7677

Experiments: Grasp Estimation Analysis

- Data from six subjects in HO3D (ABF, BB, GPMF, GSF, MDF and ShSu)
- Trained on five subjects and tested on the sixth not seen during training
- Results averaged for the six subjects
- Metric is the ADD score [1]: average distance between all vertices of a 3D mesh when transformed by the prediction compared to ground truth

[1] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, “Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes,” ACCV, 2013, pp. 548–562

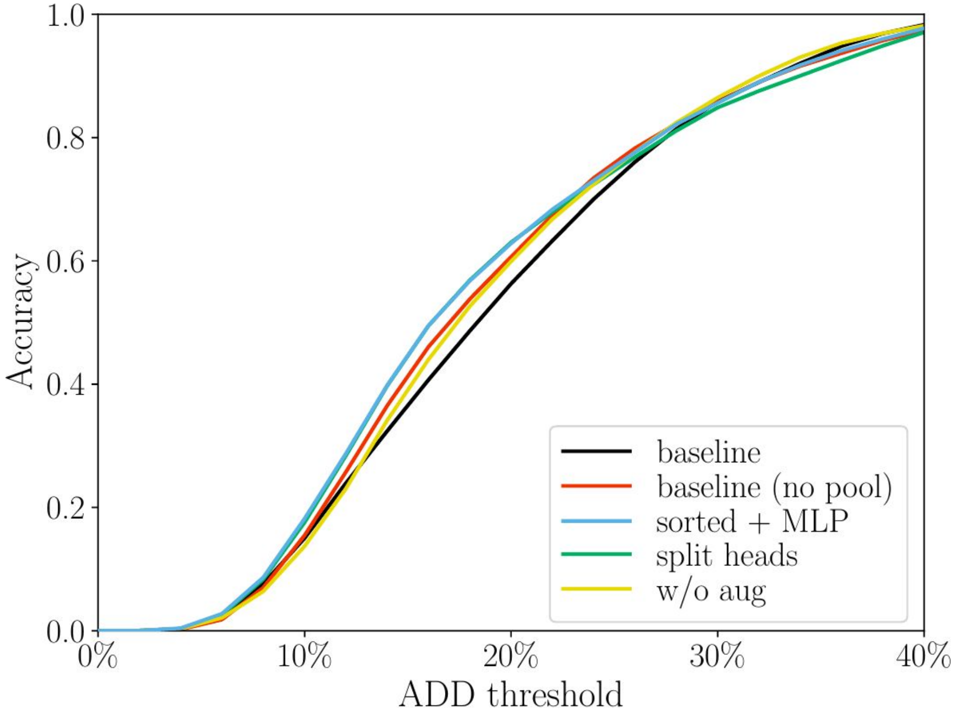
Experiments: Grasp Estimation Analysis

- Best performance when joints concatenated into high-dim. input and processed by one MLP

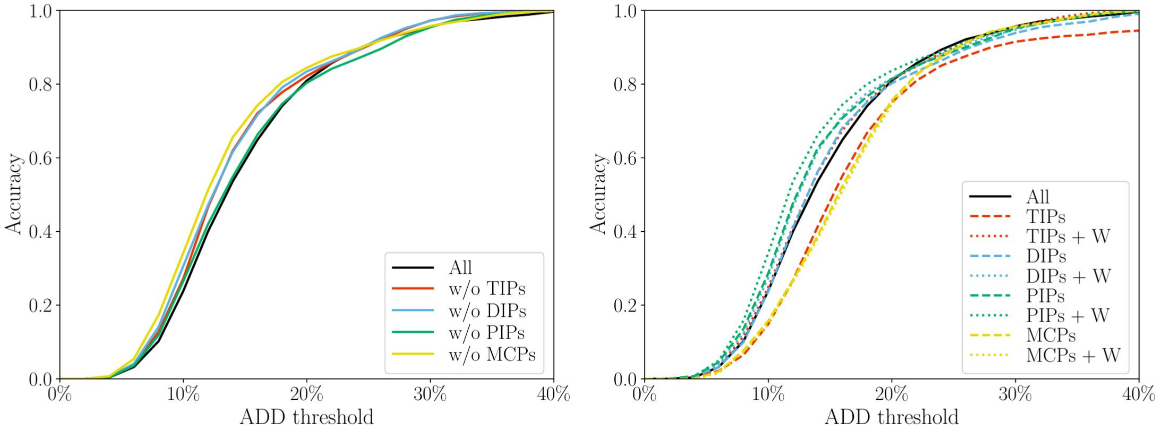
- Data augmentation provides performance boost

- Removing pooling in baseline leads to improvement

- Splitting the heads for translation and rotation does not improve accuracy



Experiments: Grasp Estimation Analysis



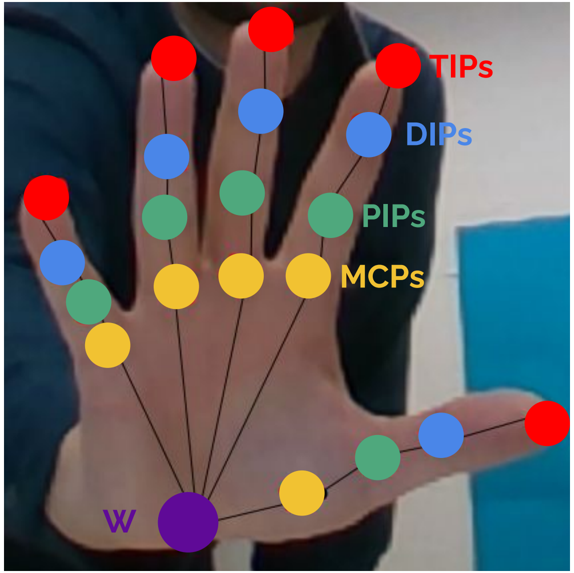


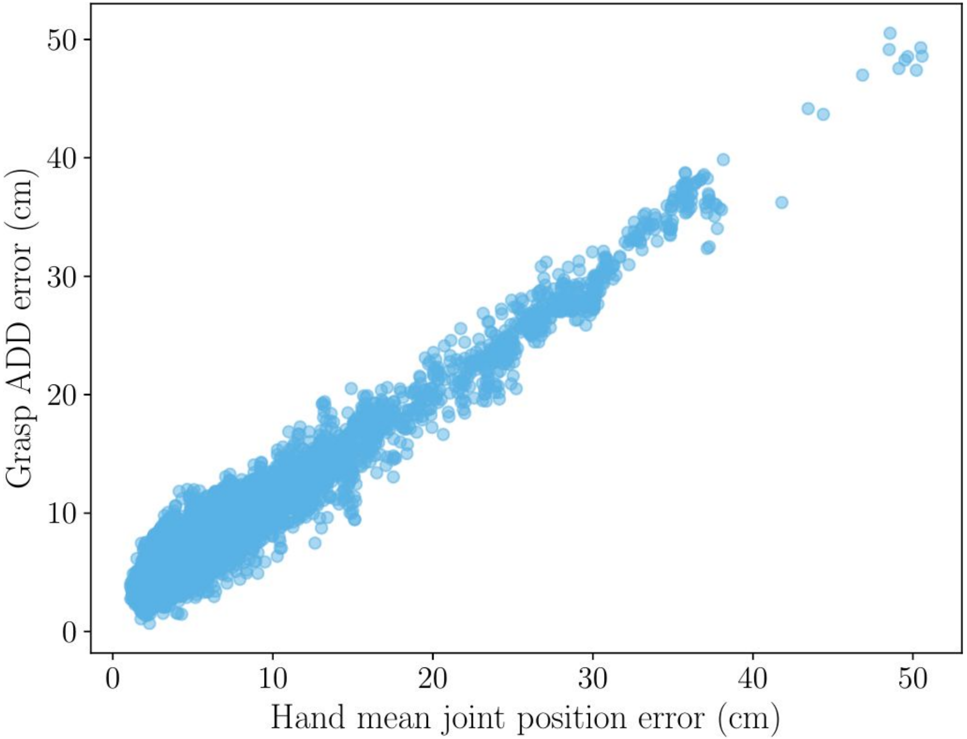
Image source: https://github.com/shreyashampali/ho3d

- Removing a single group of joints can increase accuracy
- Performance is good when using only DIPs or PIPs
- Including the wrist generally leads to improvement

Experiments: Grasp Estimation Analysis

- Hand pose estimated using tracking from [1]

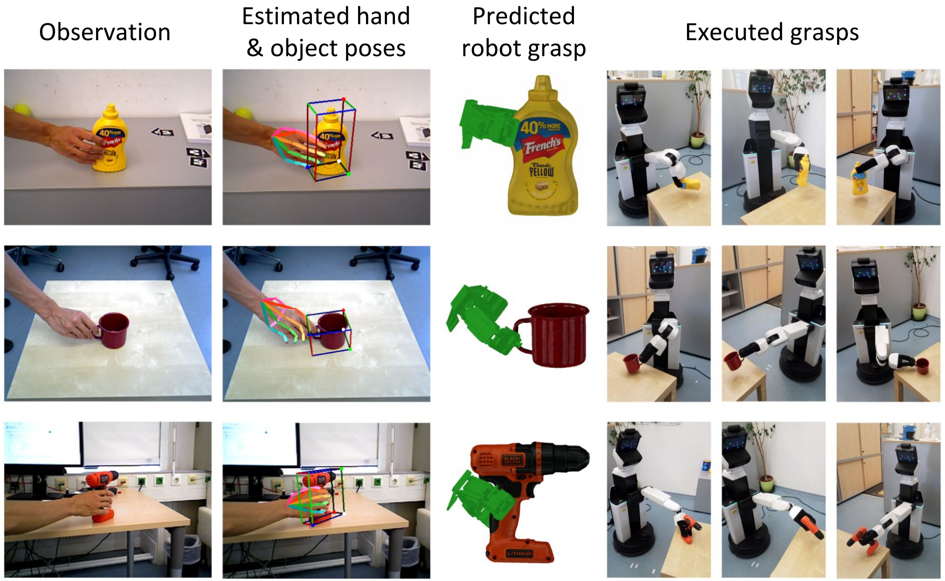
- Grasp error is correlated to hand pose error

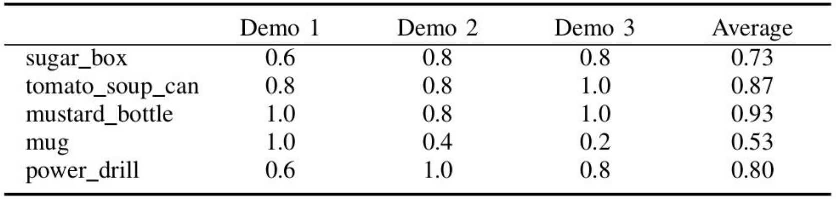


[1] P. Panteleris, I. Oikonomidis, and A. Argyros, “Using a single RGB frame for real time 3D hand pose estimation in the wild,” IEEE WACV, 2018, pp. 436–445

Experiments: Real-world Grasp Imitation

- Demonstration of full pipeline with YCB objects
- Robot successfully grasps object in new pose after observing one demonstration
- Quantitative results show good success rate for different objects
    - except for the mug, which is difficult to grasp due to thin handle

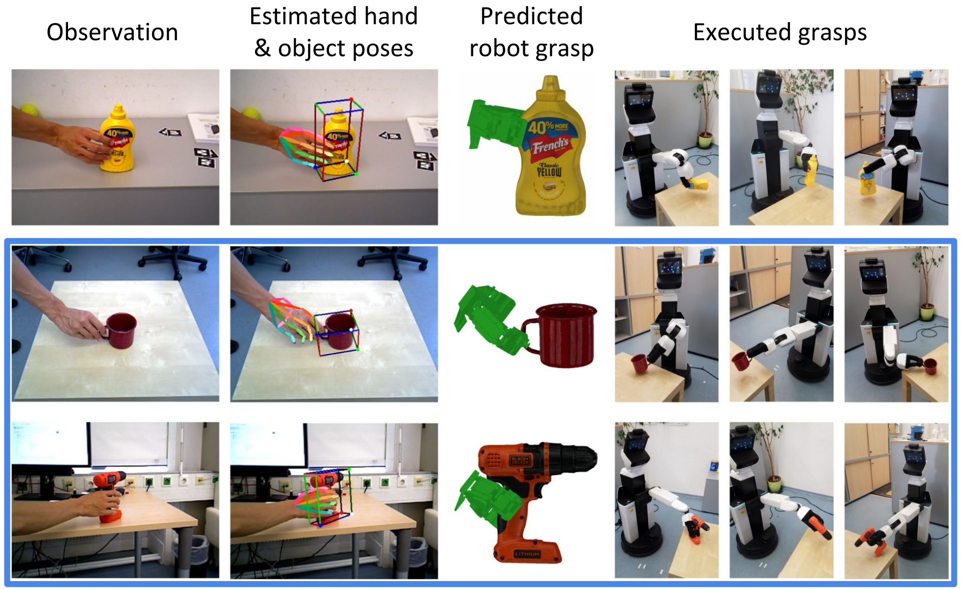


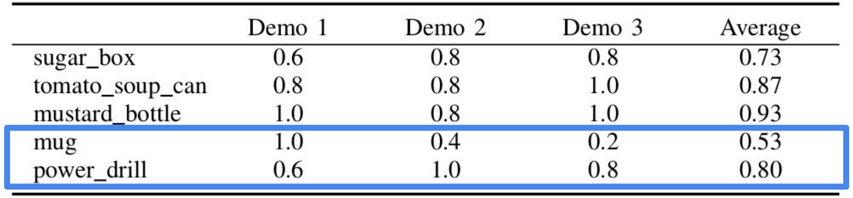


	Demo 1	Demo 2	Demo 3	Average
sugar_box	0.6	0.8	0.8	0.73
tomato_soup_can	0.8	0.8	1.0	0.87
mustard_bottle	1.0	0.8	1.0	0.93
mug	1.0	0.4	0.2	0.53
power_drill	0.6	1.0	0.8	0.80

Experiments: Real-world Grasp Imitation

- Demonstration of full pipeline with YCB objects
- Robot successfully grasps object in new pose after observing one demonstration
- Quantitative results show good success rate for different objects
  - except for the mug, which is difficult to grasp due to thin handle
- Task-oriented grasps for objects with handles





	Demo 1	Demo 2	Demo 3	Average
sugar_box	0.6	0.8	0.8	0.73
tomato_soup_can	0.8	0.8	1.0	0.87
mustard_bottle	1.0	0.8	1.0	0.93
mug	1.0	0.4	0.2	0.53
power_drill	0.6	1.0	0.8	0.80

Thank you

The research leading to these results has received funding from the Austrian Science Fund (FWF) under grant agreement No. I3969-N30 (InDex)

ACIN TU WIEN

IARIA The Sixteenth International Conference on Autonomic and Autonomous Systems
ICAS 2020
September 27, 2020 to October 01, 2020 - Lisbon, Portugal