Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Learning to Grasp through Self-Supervision

Lukas Sommerhalder

14.12.2023

Picture: Murali et al.

Agenda

1. Introduction
2. Examples
3. Comparison
4. Demonstration
5. Questions & Discussion

1

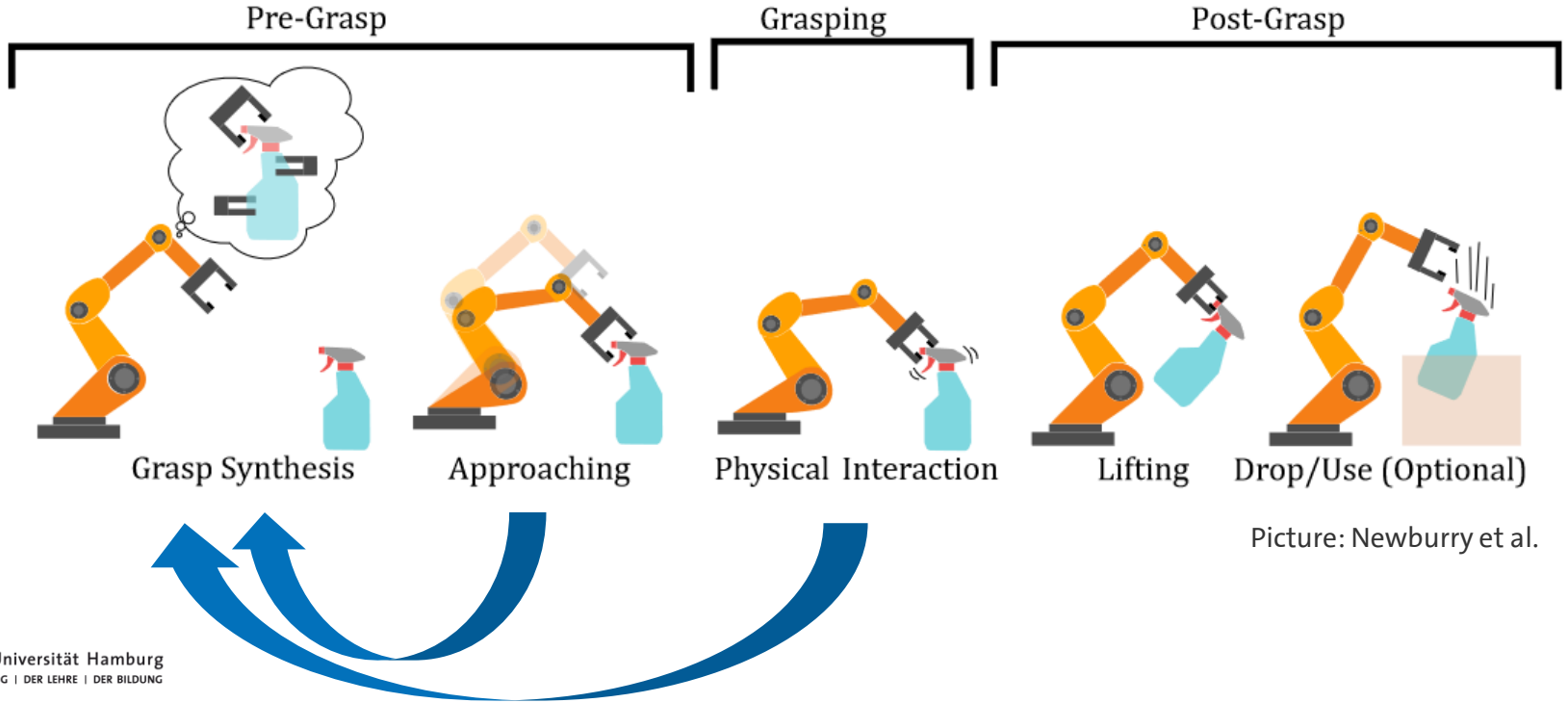
Introduction

Photo: Pinto and Gupta

Goals

- Provide answers/insights to the following:
  - What is a Grasp/Grasp-Process?
  - What are the benefits of self-supervised learning for grasping?
  - How can we achieve successful grasps through self-supervised learning?
  - An alternative approach that is neither supervised nor unsupervised.
- Create an intuition to solve such and similar problems

Grasp Process



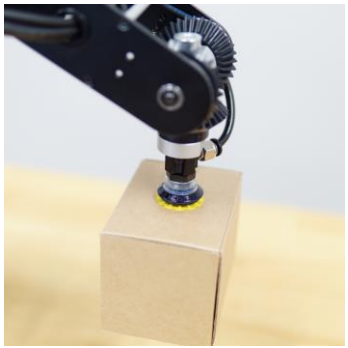
Picture: Newburry et al.

Tools to Grasp

2-Jaw Gripper

3-Finger Gripper

Vacuum Gripper



Bionic Hand

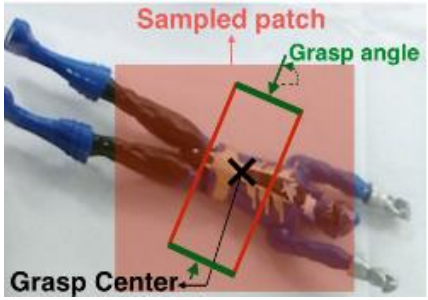
Tools used for Research

- Literature Review from Newburry et al., 2022, about grasp synthesis with Deep Learning (6DoF), 85 reviewed papers.
  - Most commonly used gripper: Two-Finger parallel jaw (51 times)
  - Most commonly used robot: Robotic arm (66 times)

Manufacturer	Model	Popularity	DoF
Robotiq	2F	10	1
Robotiq	3F	5	1
Franka Emika	Panda Gripper	12	1
Barret	BarretHand	7	5
Kinova	2F	3	1
Kinova	3F	1	2
Wonik Robotics	Allegro Hand	4	16
Rethink Robotics	Baxter Gripper	4	1
Shadow Robot Company	Shadow Hand	2	20
	Custom	2	
	Other	32	

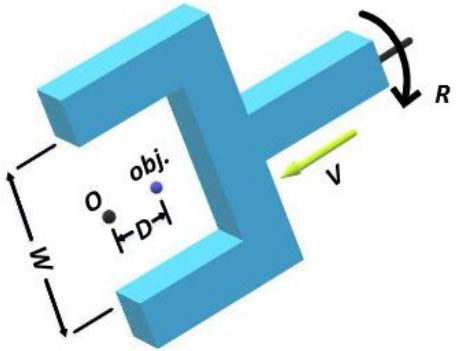
Input Format	Number of times used
Point Cloud	34
Depth Image	18
RGB-D Image	15
Voxel Grid	12
Segmentation Mask	9
Other	10

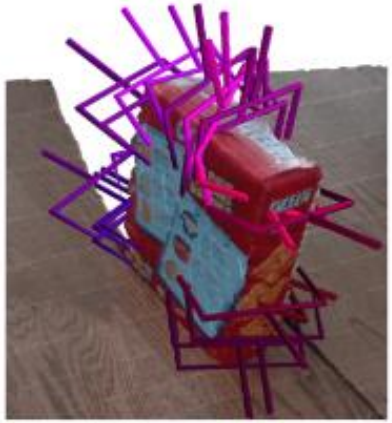
Grasp Pose



3-DoF

7-DoF





6/7-DoF

Methods for Grasp Synthesis

- **Analytical Methods**
  - Requires an accurate model of the object (hard to get)
- **Supervised / Human labeling**
  - Biased
  - Human working time
- **Self-Supervised Learning**
  - Autonomous label generation/learning (without human bias/effort)
  - Labeling through Trial-and-Error
  - Simulation / Roboter working time
- **Reinforcement Learning**

2

Examples

Photo: S. Levine et al.

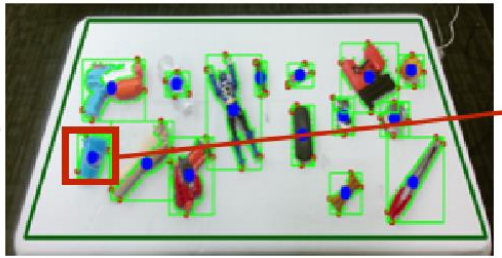
Three different Approaches

1. **Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours.**
   - L. **Pinto** and A. **Gupta**, 2015, Carnegie Mellon University.
2. **Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection.**
   - S. **Levine**, P. **Pastor**, A. **Kryzhevsky** and D. **Quillen**, 2017, Google.
3. **GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping**
   - H.-S. **Fang**, C. **Wang**, M. **Gou**. C. **Lu**, 2020, Shanghai Jiao Tong University.

Supersizing Self-Supervision

- Pinto and Gupta (2015), Carnegie Mellon University
  - RGB-Camera (known Position)
  - Pretrained Convolutional Layers (AlexNet trained on ImageNet)
  - Random attempts and evaluation
  - Classification into 18 different angles, 3-DOF annotation

Query Kinect image



Find objects via MOG subtraction

Approach random object

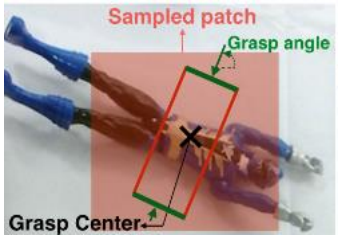
Execute random grasp

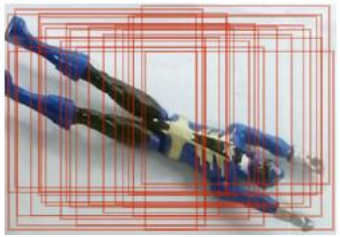
Verify grasp success

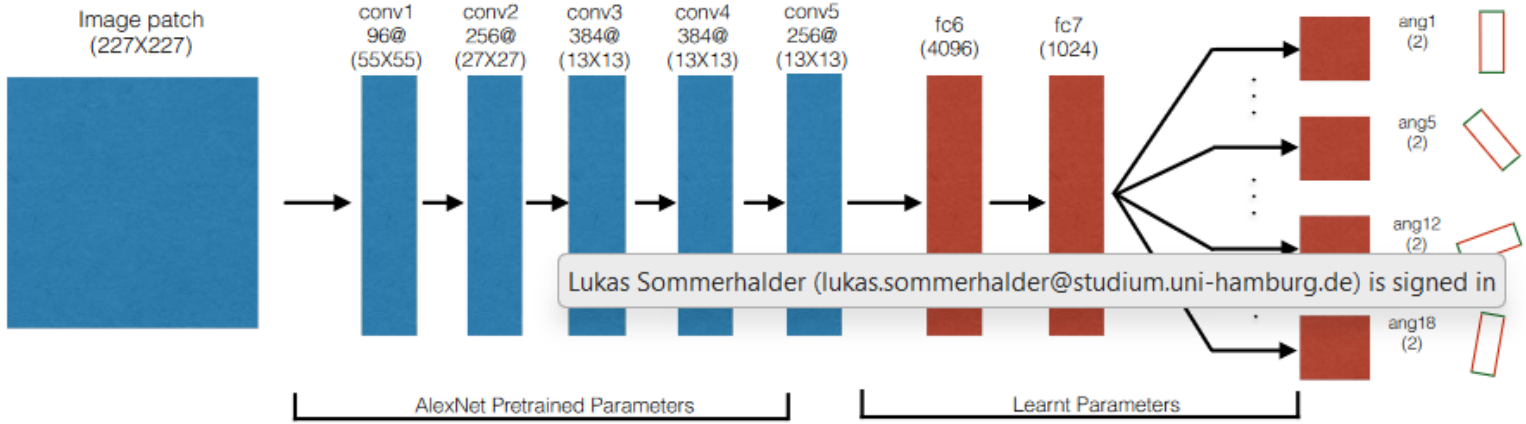
Supersizing Self-Supervision: CNN

The Convolutional Neural Network (CNN) Architecture

- Input: Image patch centered at the presumed grasp point
- Output: 18 scores







Supersizing Self-Supervision: Dataset

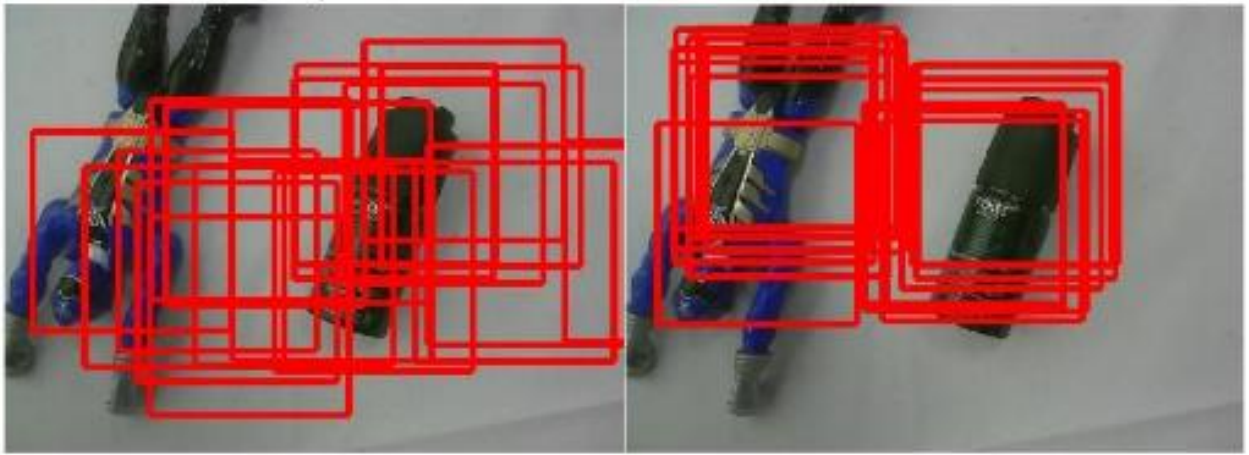
- 50k+ grasp attempts
  - Random 40k for the training set and 3k for the test set
  - 7k with pretrained network

Data Collection Type	Positive	Negative	Total	Grasp Rate
Random Trials	3,245	37,042	40,287	8.05%
Multi-Staged	2,807	4,500	7,307	38.41%
Test Set	214	2,759	2,973	7.19%
	6,266	**44,301**	**50,567**	

Supersizing Self-Supervision: Staged Learning

Staged Learning

1. Trained classification CNN-Model as prior
2. Previous stage with seen and unseen objects



Supersizing Self-Supervision: Learning/Dataset

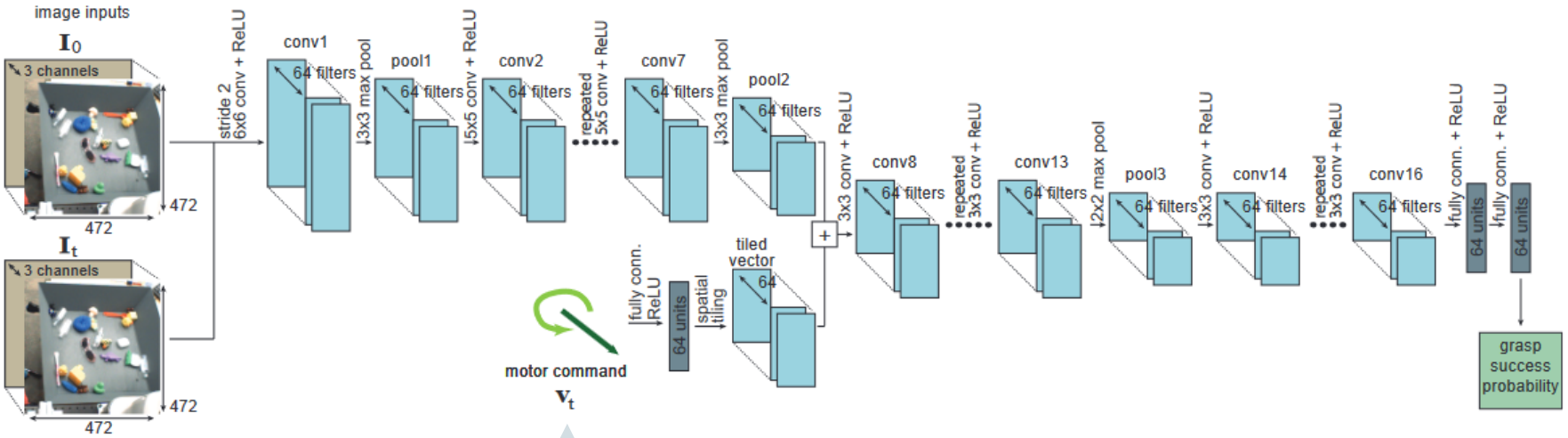
Binary classification of the test set (single stage)

The success rate is 66% for unseen and 73% for seen objects

Hand-Eye Coordination

- Levine et al. (2016), Google
  - RGB-Camera (unknown position)
  - Tries to learn the best next move to a grasp
  - „Can react to changes during the grasp process“

Hand-Eye Coordination: CNN



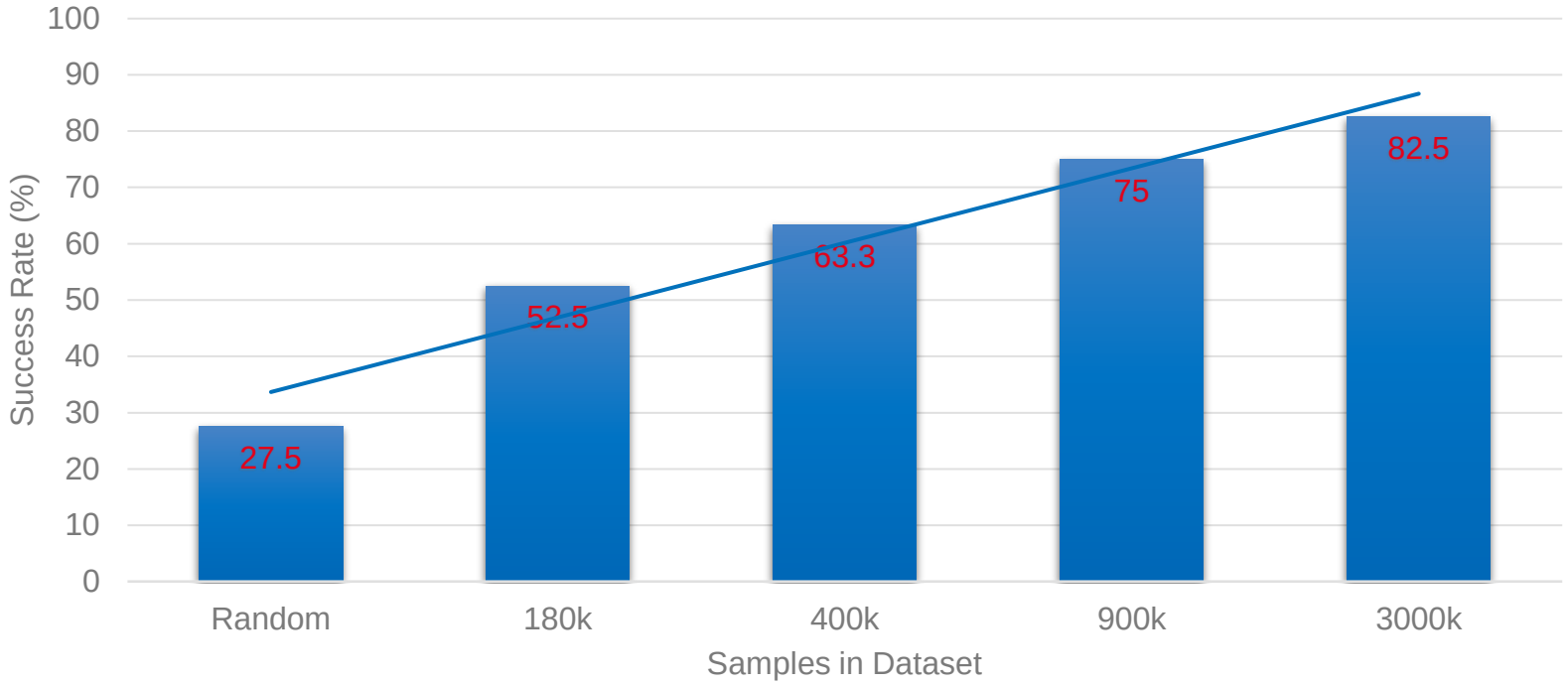
$v_t = p_{t+1} - p_t$

Hand-Eye Coordination: Learning Process

Staged Learning

1. Start with random motor commands (half of the data), single step
2. Trained model from the last stage
   - Cross Entropy Method (CEM) to find the next move
   - Projection of the „best“ next move to the table height
3. Increasing max steps by two (till 10), and train again

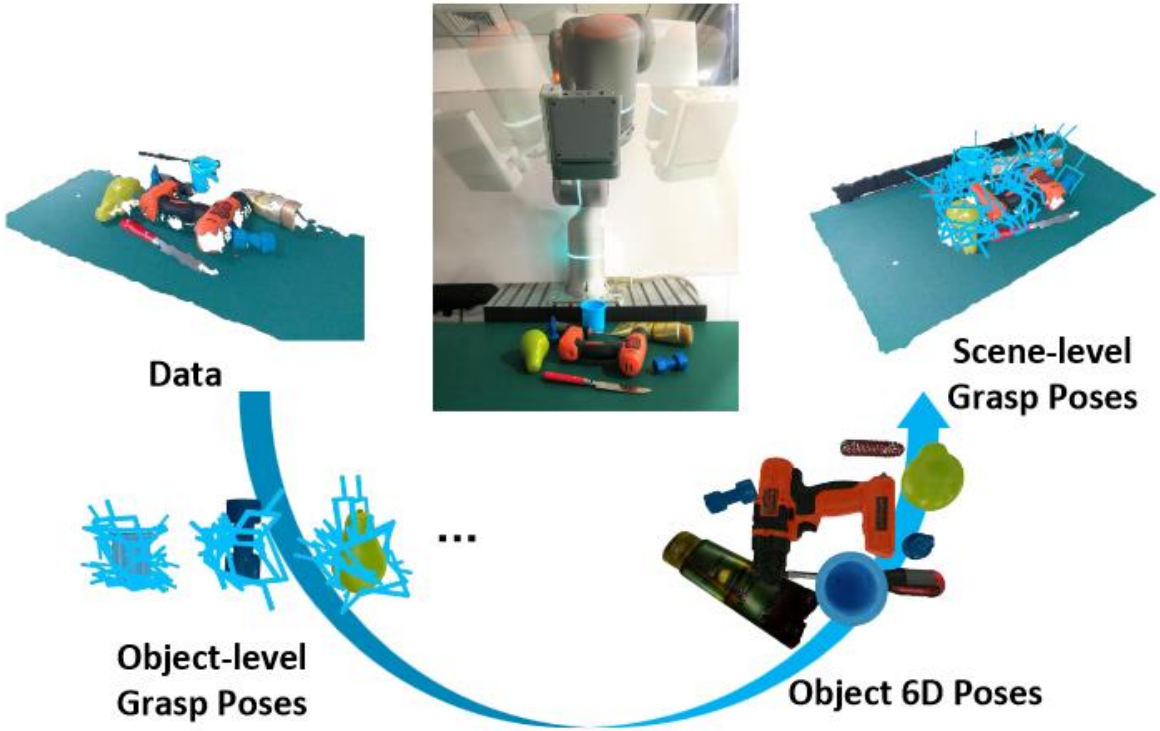
Hand-Eye Coordination: Learning/Dataset



GraspNet-1Billion

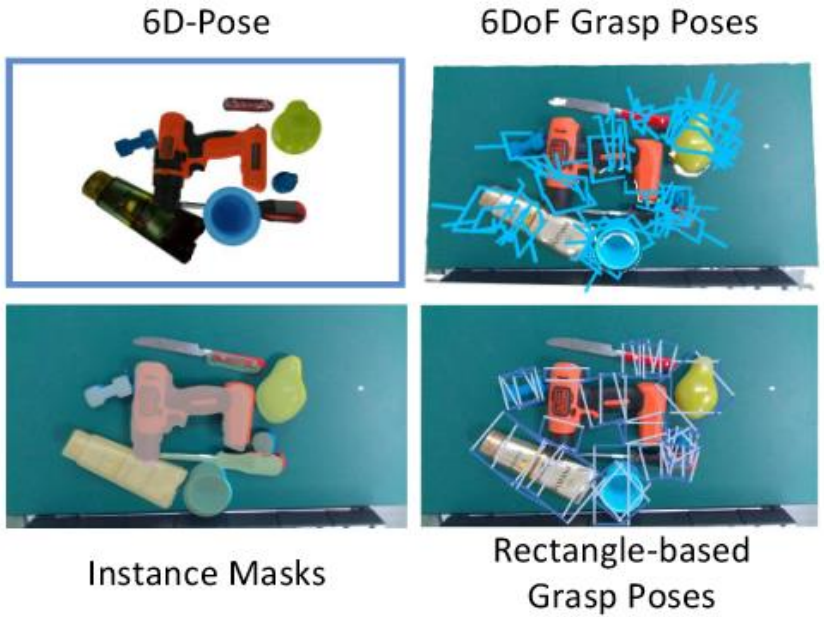
- Fang et al. (2020), Shanghai Jiao Tong University
- Grasp annotation of real-world data through analytical computation on sim objects
  - Depth cloud (RGB-D camera) (multiple viewpoints)
  - The DNN tries to adapt depth clouds from real objects to eighty-eight known 3D-Models
  - Capable of 7-DoF grasp synthesis and collision detection

GraspNet-1Billion: Grasp/Pose Synthesis

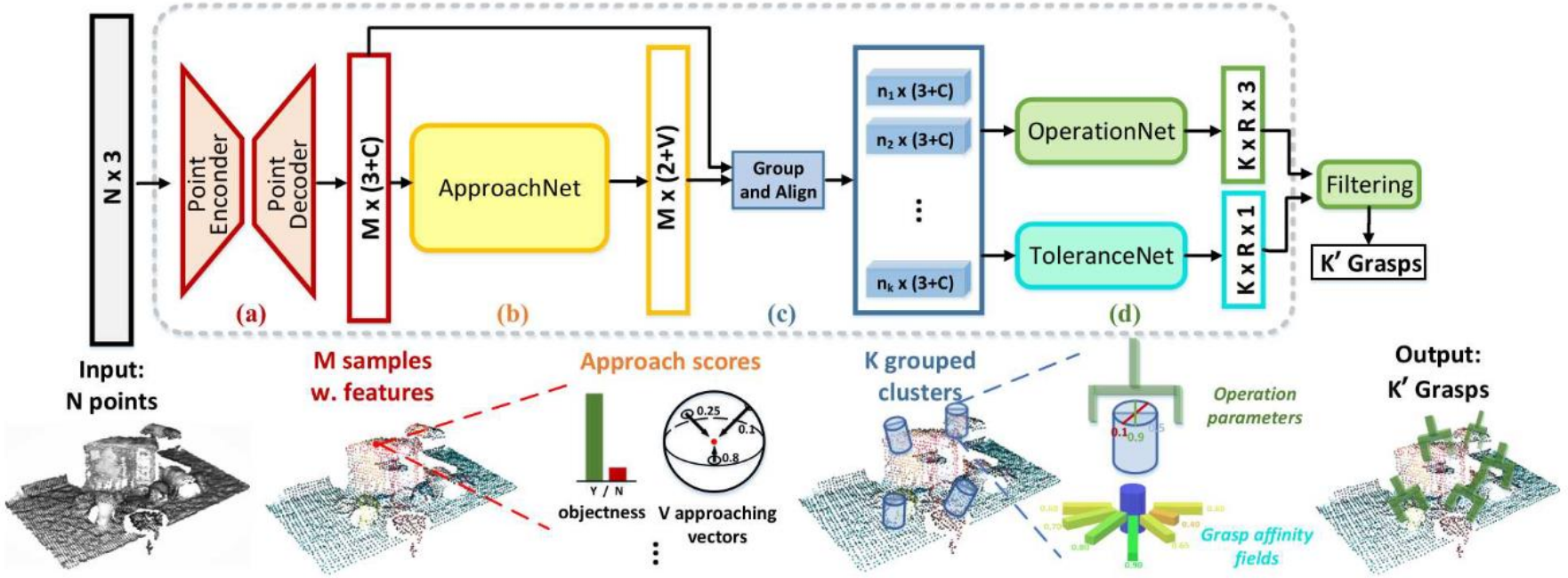


GraspNet-1Billion: Dataset

- **Dataset**
  - 97k Images from 190 cluttered scenes
  - 88 objects
  - Leads to more than 1 Billion 6-DoF grasp annotations



GraspNet-1Billion: DNN



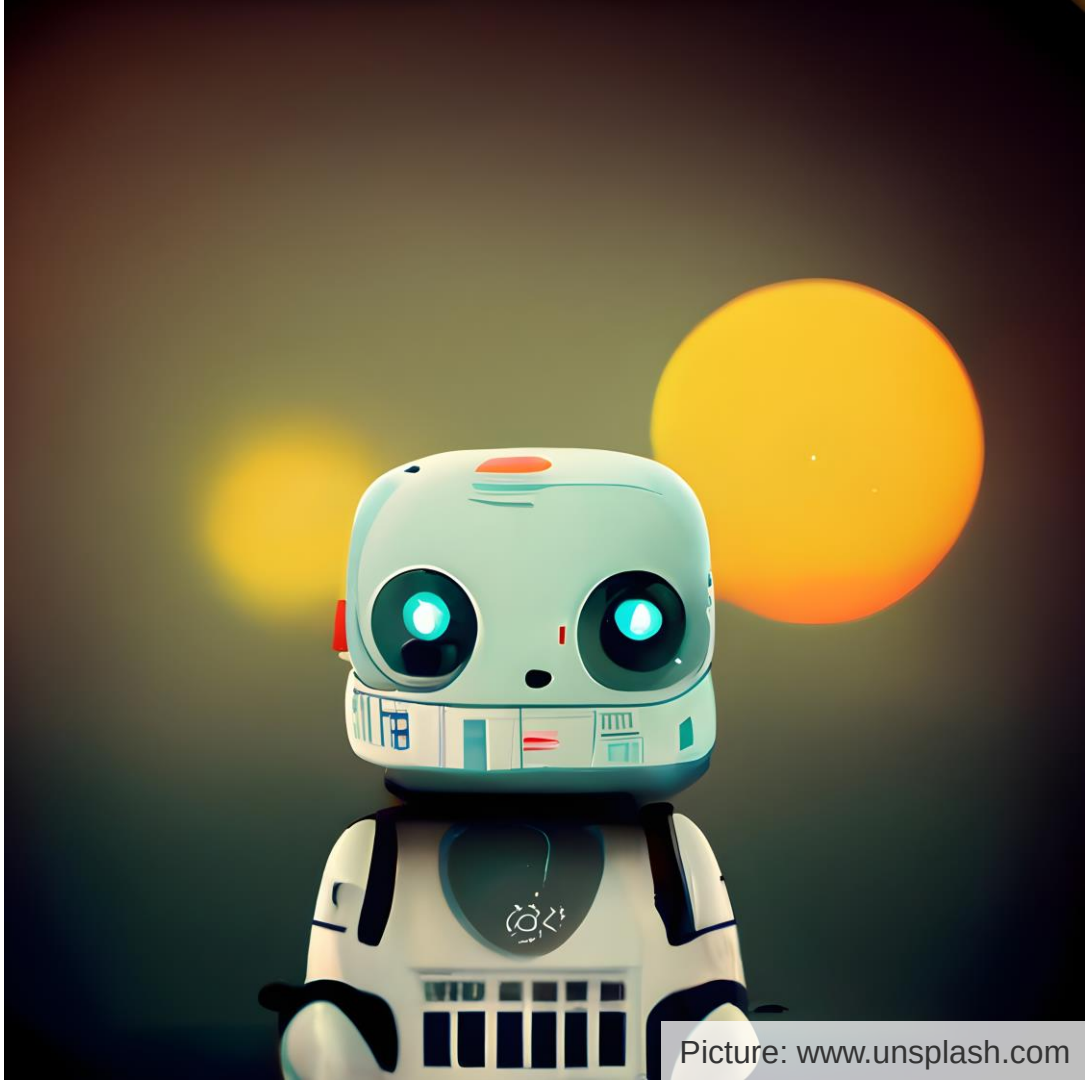
GraspNet-1Billion: Ablation Study

Comparison of grasp score (s) and success rate

Object	s=1	s=0.5	s=0.1	Object	s=1	s=0.5	s=0.1
Banana	98%	67%	21%	Apple	97%	65%	16%
Peeler	95%	59%	9%	Dragon	96%	60%	9%
Mug	96%	62%	12%	Camel	93%	67%	23%
Scissors	89%	61%	5%	Power Drill	96%	61%	14%
Lion	98%	68%	16%	Black Mouse	98%	64%	13%

3

Conclusion



Picture: www.unsplash.com

Comparison

	50k Tries	Eye-Hand Coordination	GraspNet
Dataset Size	50k	3000k	1000k
Sensor	RGB-Camera	RGB-Camera	RGBD-Camera
Main Feature	Simple CNN	Motion Planning Risk of collisions	7-DoF Pose Big data set

(No direct comparison of performance due to different setups, objects and goal parameters)

Conclusion

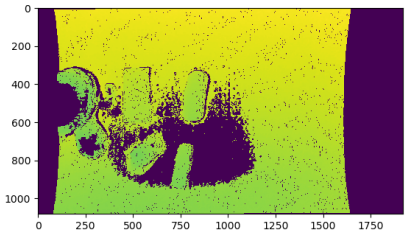
- Importance of Large-Scale Data and Multi-Stage learning
- Improving grasping models in robotics
- There are various approaches to solve the same problem

4

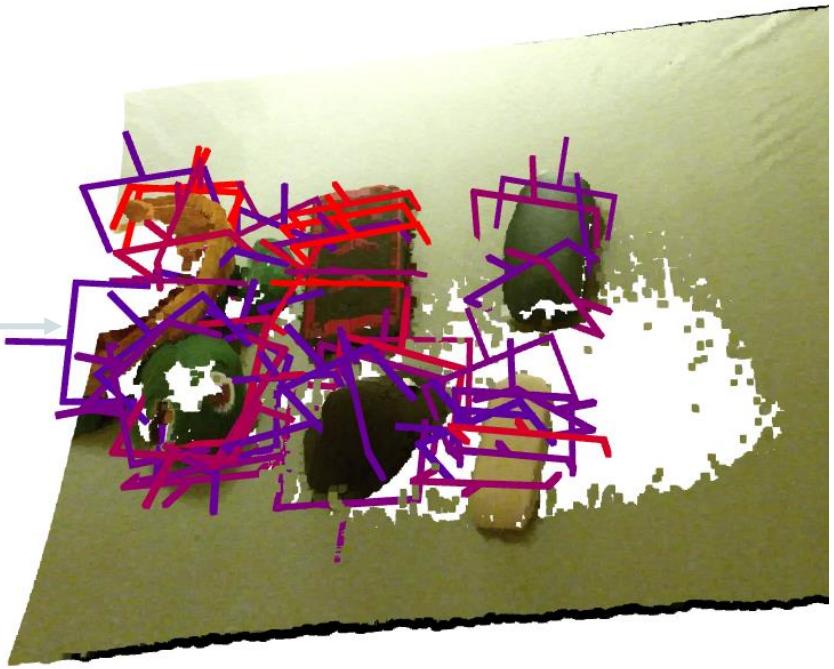
Demonstration

Picture: GraspNet

GraspNet: 7-DoF Grasp Pose Synthesis



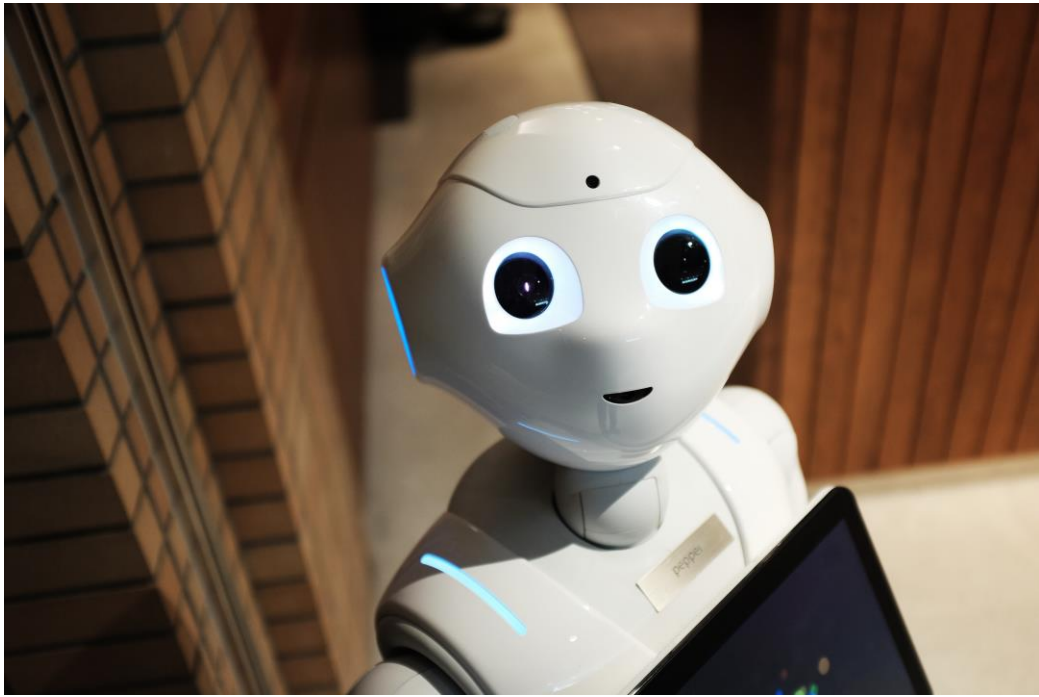
PC



Sourcecode: https://github.com/graspnet/graspnet-baseline

5

Questions & Discussion



Picture: www.unsplash.com

References

- H.-S. Fang, C. Wang, M. Gou. and C. Lu, GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 11441-11450. https://doi.org/10.1109/CVPR42600.2020.01146 .

- S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research. 2018, 37(4-5), pp. 421-436. https://doi.org/10.1177/0278364917710318.

- R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, A. Cosgun, Deep Learning Approaches to Grasp Synthesis: A Review. 2022, Australian National University Canberra. https://doi.org/10.48550/arXiv.2207.02556.

- L. Pinto and A. Gupta, Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours. *2016 IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2016, pp. 3406-3413. https://doi.org/10.1109/ICRA.2016.7487517.

A. Deep Learning Approaches to Grasp Synthesis: A Review

- **Inclusion/Exclusion Criteria of Papers**
  - From the papers found in the databases, only publications that met the following criteria are included in this review:
    - Paper considered grasping from a table-top scenario,
    - All 6-DoF were used for the grasp pose,
    - Deep Learning methods were applied in some aspect of the work,
    - Published after Jan 1, 2012, (the year AlexNet was published) and Written in English

B. Hand-Eye coordination

- **Cross Entropy Method**
  - 64 samples and 3 iterations with the 6 best grasp directions, starting with a normal distribution at the current position.
  - All samples are constrained to the workspace and rotations between
  - All grasp directions are projected to the table height.