Branch Dueling Deep Q-Networks for Robotics Applications

Master's Thesis

Baris Yazici

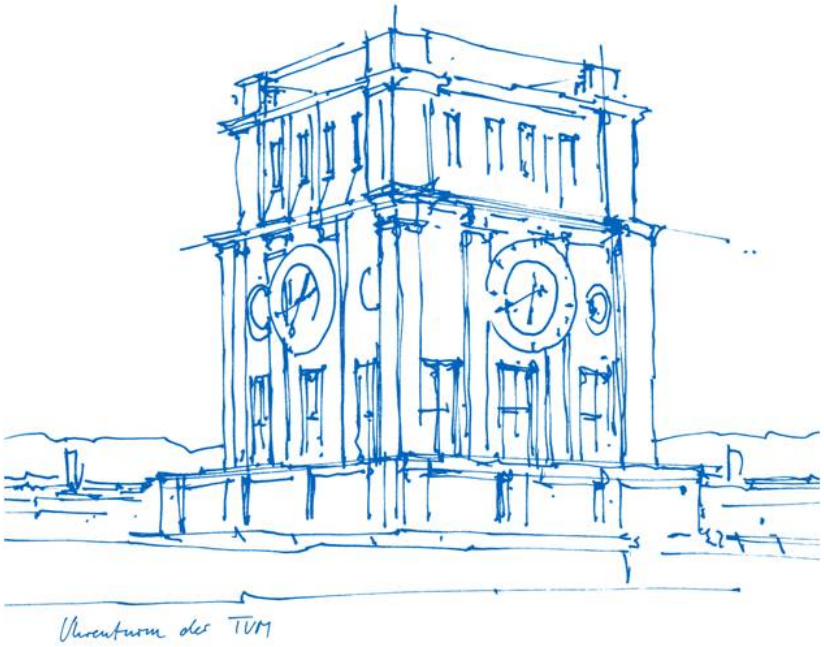
Msc. Mahmoud Akl

Technical University of Munich

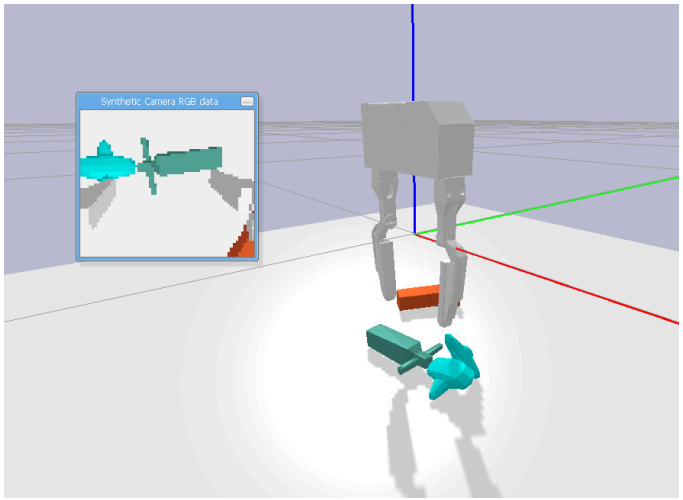
Computer Science Faculty

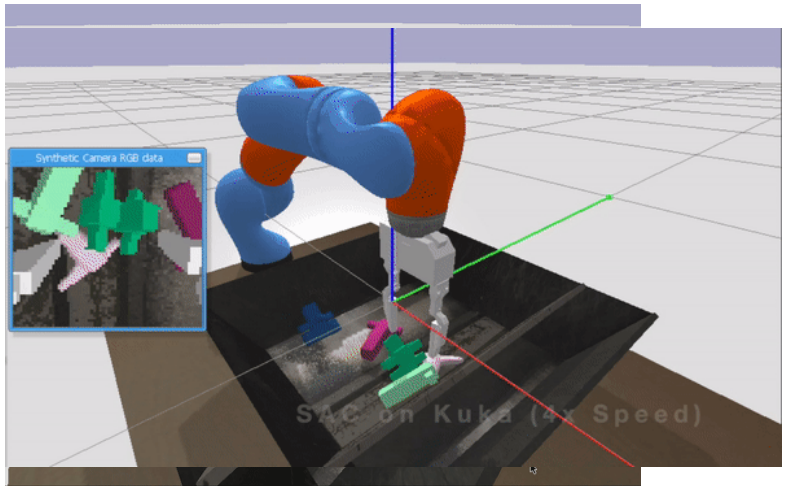
Chair for Robotics, Artificial Intelligence and Realtime Systems

Munich, 05.10.2020

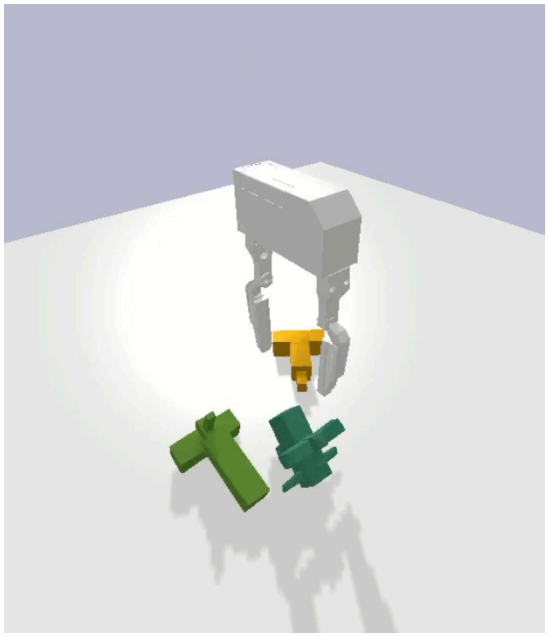


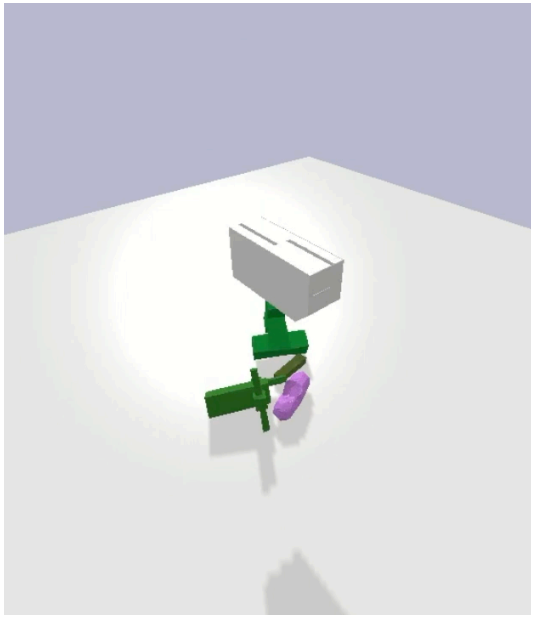
Adaptive Grasping with Domain Transfer



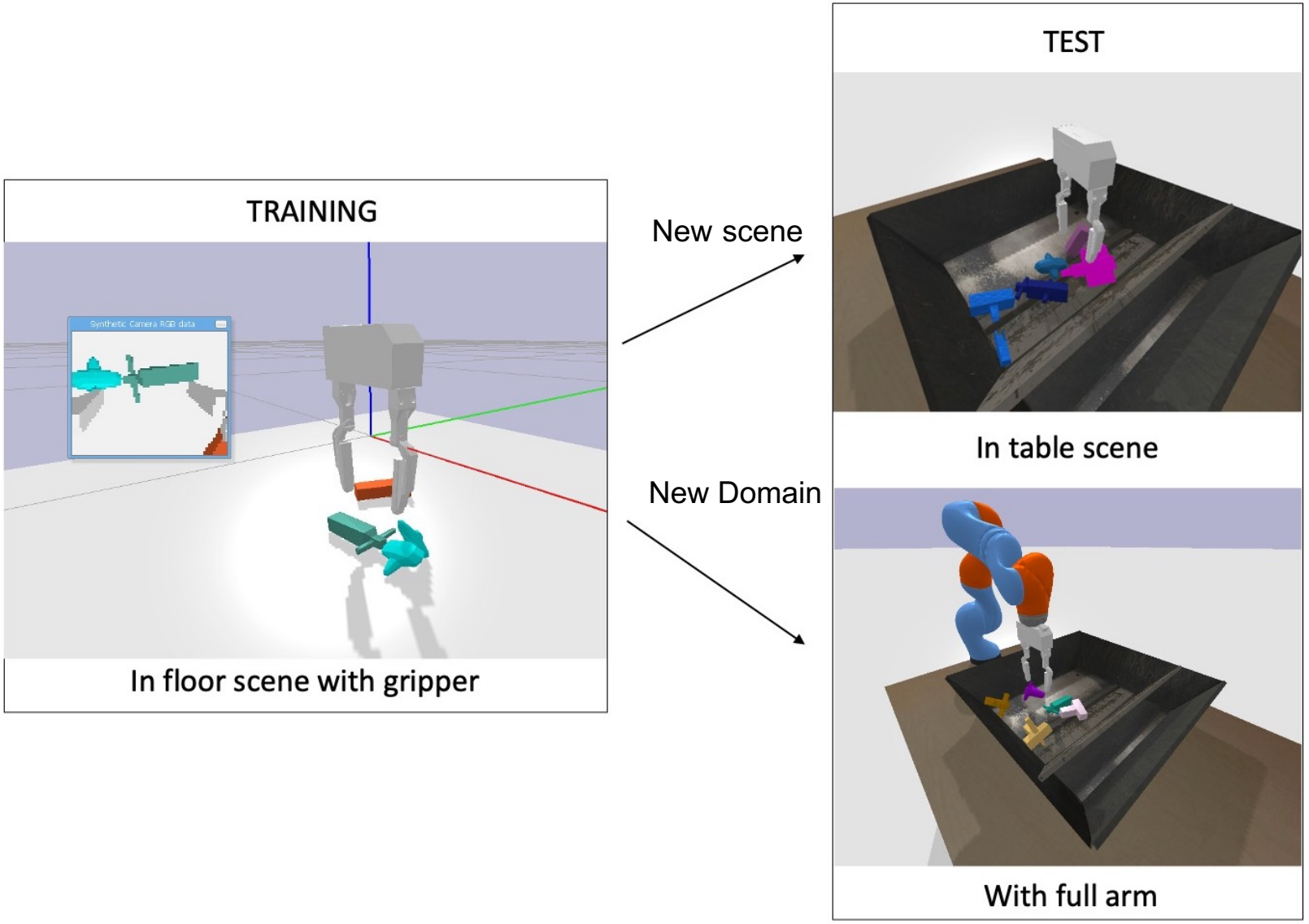


Train one grasping model for different **scenes** and **environments**

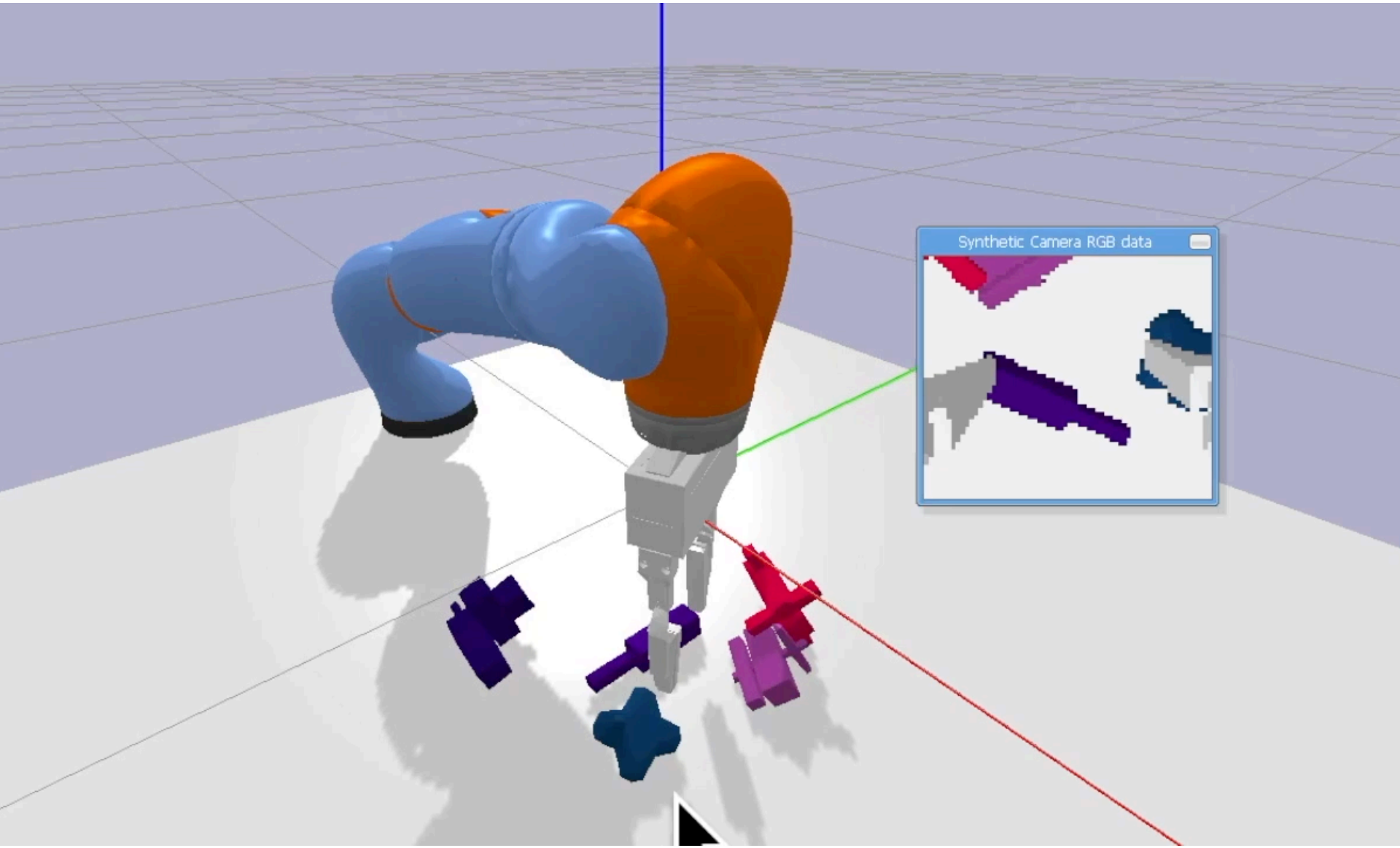




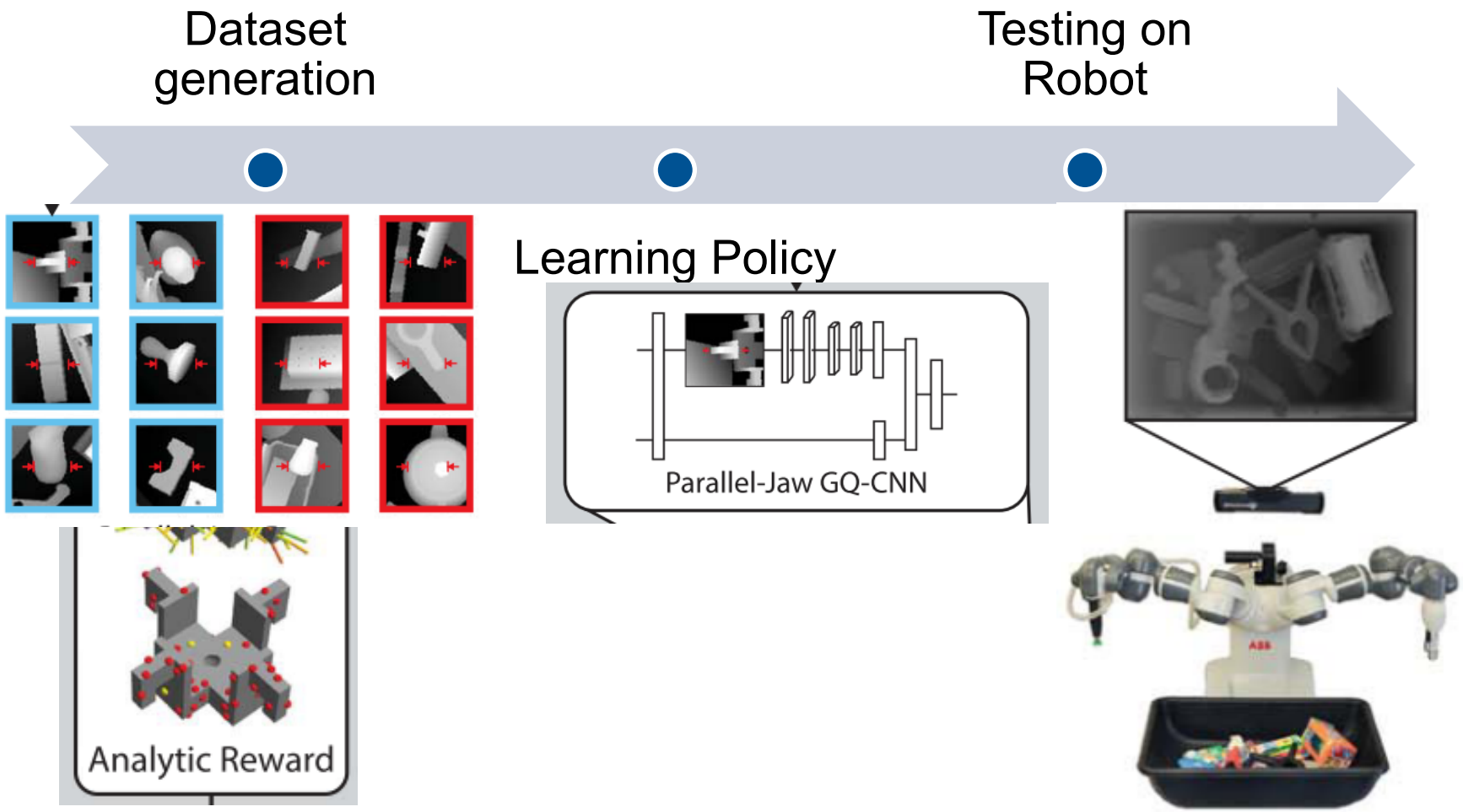
Train-Test Pipeline



Why Adaptive Robust Grasping?



Current State of Art



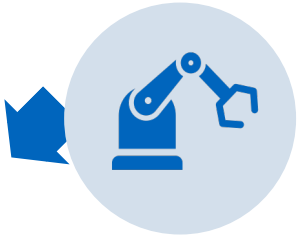
Mahler et al., 2019

Our Approach

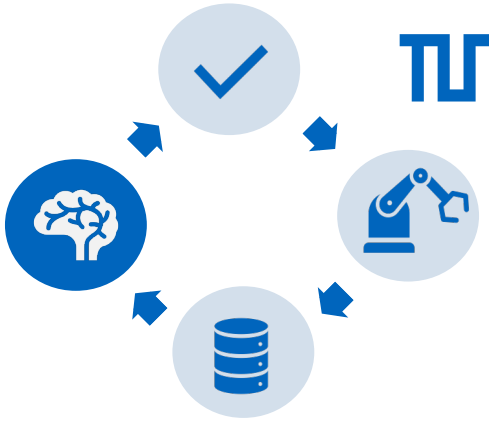
Learn Policy

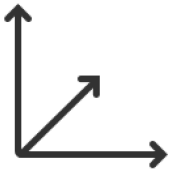
Select Action

Feed Data to Buffer



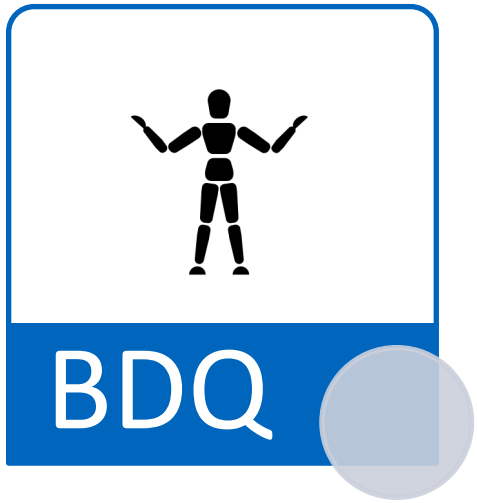
Act On Robot

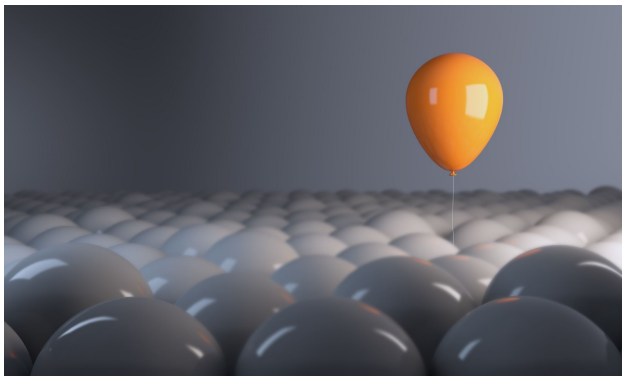


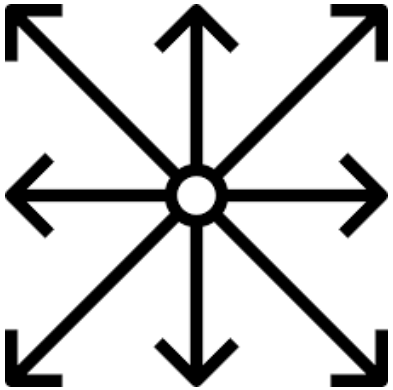


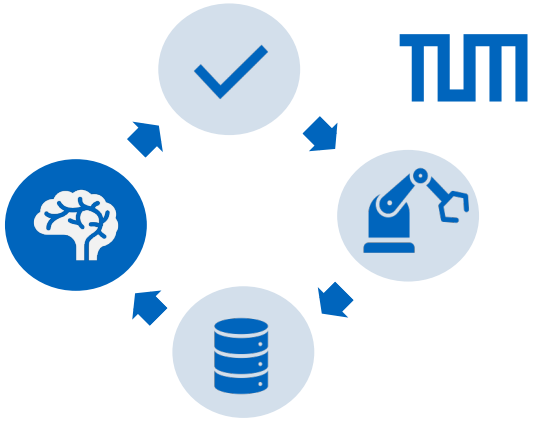

SAC

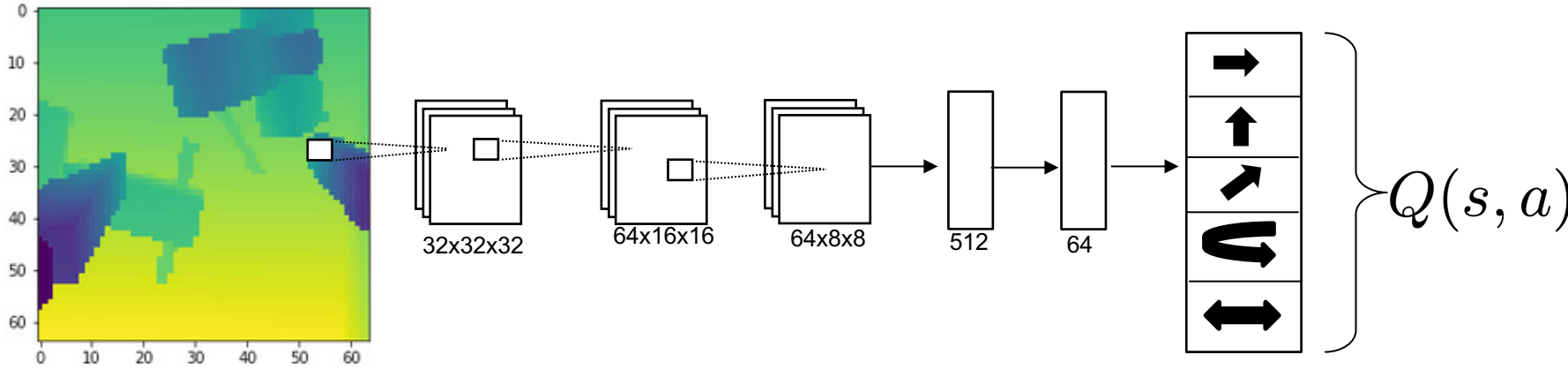

DQN


BDQ





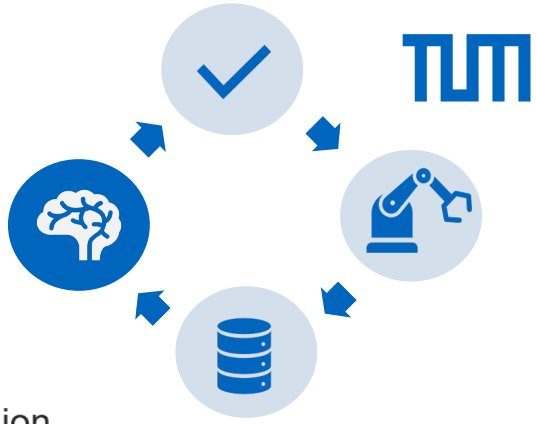


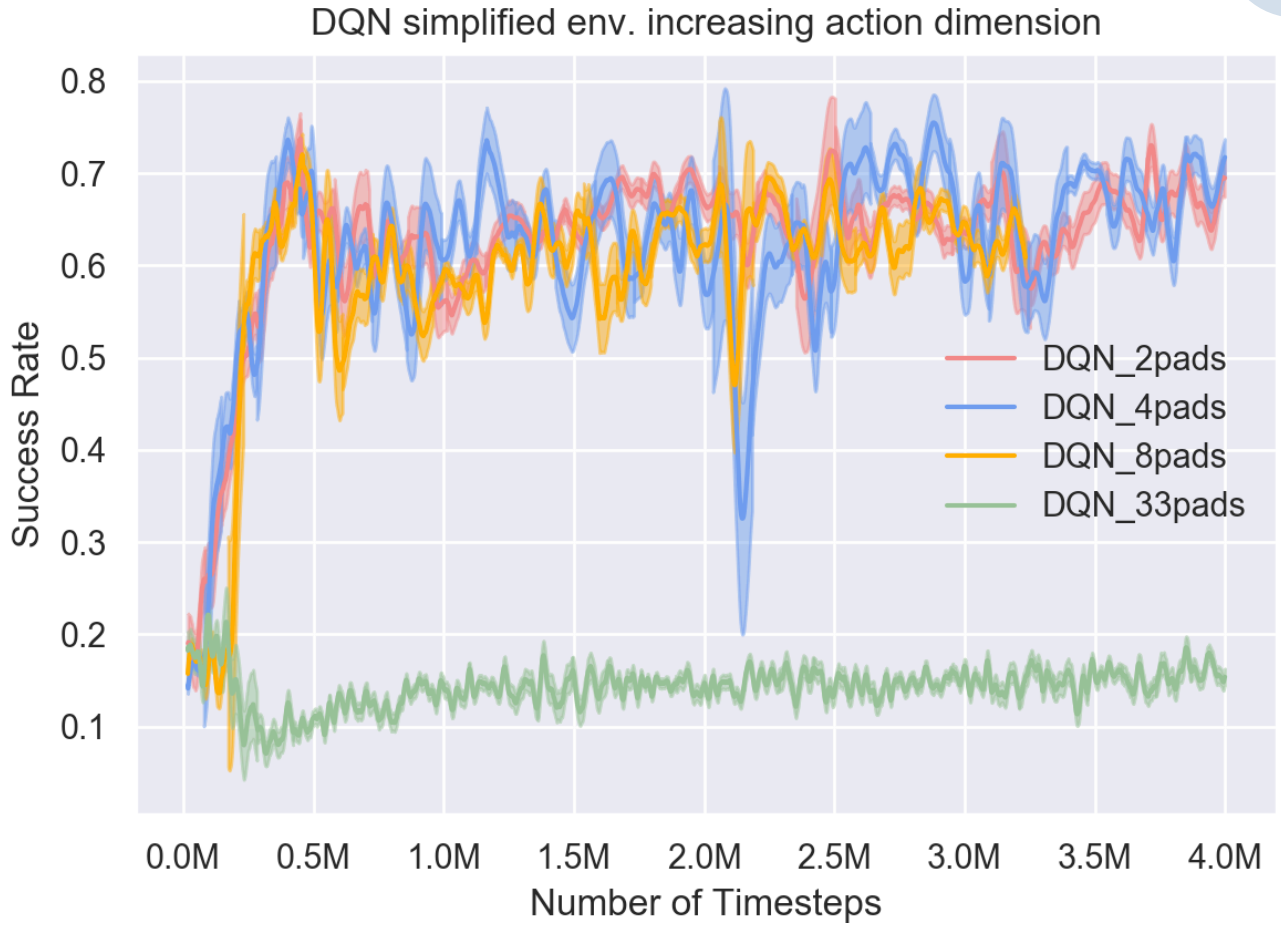


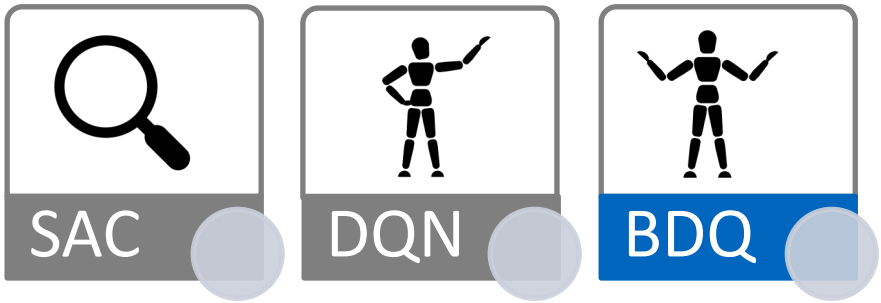
$$L(\theta) = E_{(s,a,r,s')\sim U(D)}\left[Q_t(s', a; \theta^-) - Q(s, a; \theta)\Big)^2\right]$$

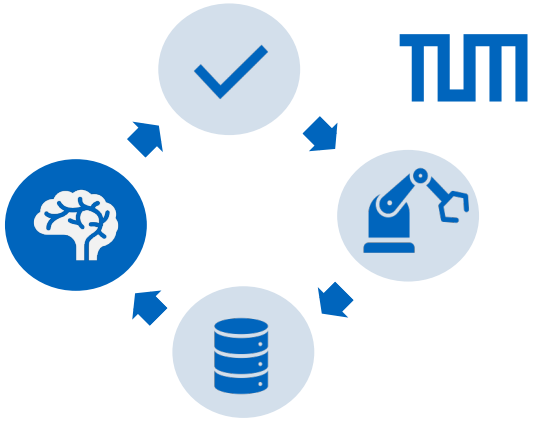
$$Q^*(s,a) = \mathbb{E}_{s'\sim\mathcal{E}}\left[r + \gamma \max_{a'} Q^*(s', a')\middle| s, a\right]$$

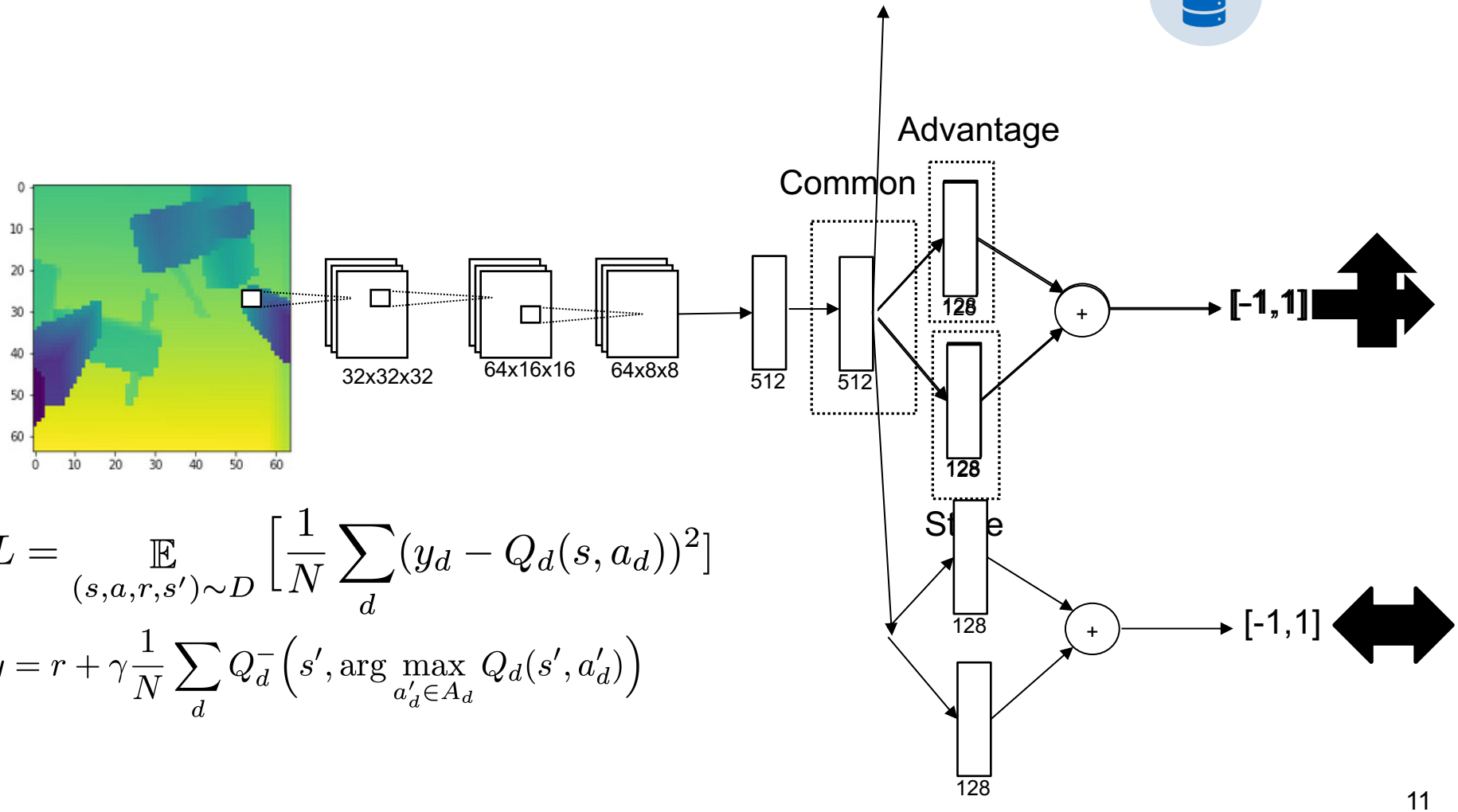

SAC
DQN
BDQ




DQN simplified env. increasing action dimension
Success Rate
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0M
0.5M
1.0M
1.5M
2.0M
2.5M
3.0M
3.5M
4.0M
Number of Timesteps
DQN_2pads
DQN_4pads
DQN_8pads
DQN_33pads

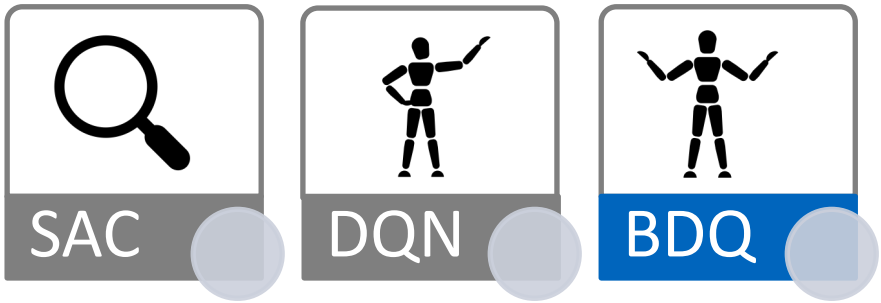


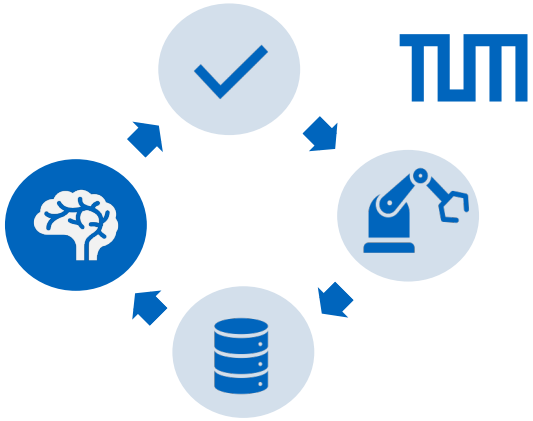


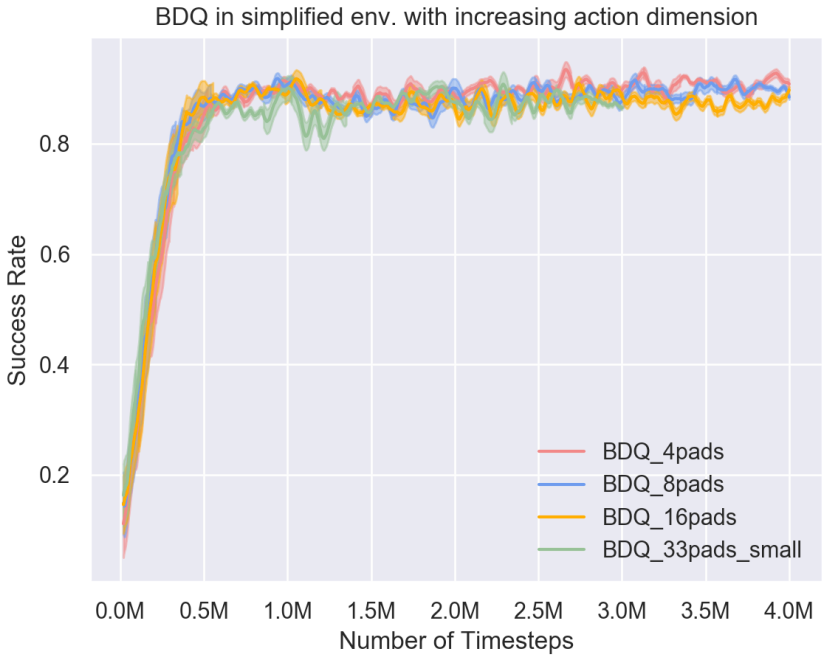


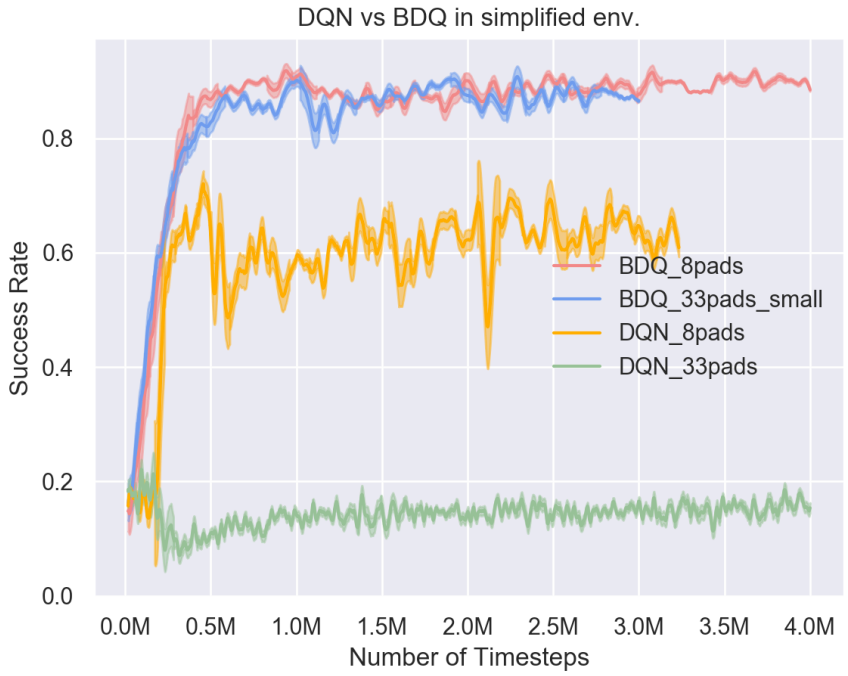
$$L = \mathop{\mathbb{E}}_{(s,a,r,s')\sim D}\left[\frac{1}{N}\sum_{d}(y_d - Q_d(s,a_d))^2\right]$$

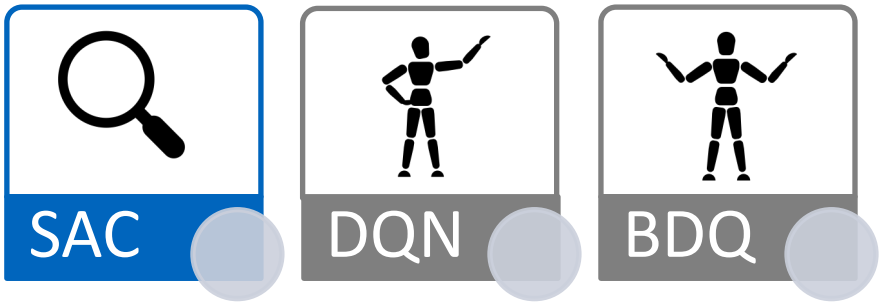
$$y = r + \gamma\frac{1}{N}\sum_{d} Q_d^{-}\left(s', \arg\max_{a'_d \in A_d} Q_d(s', a'_d)\right)$$

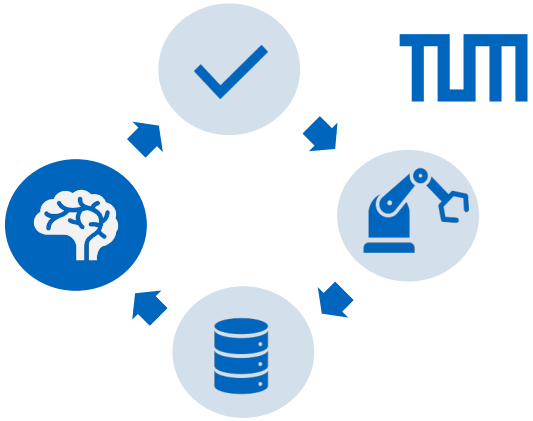

SAC
DQN
BDQ

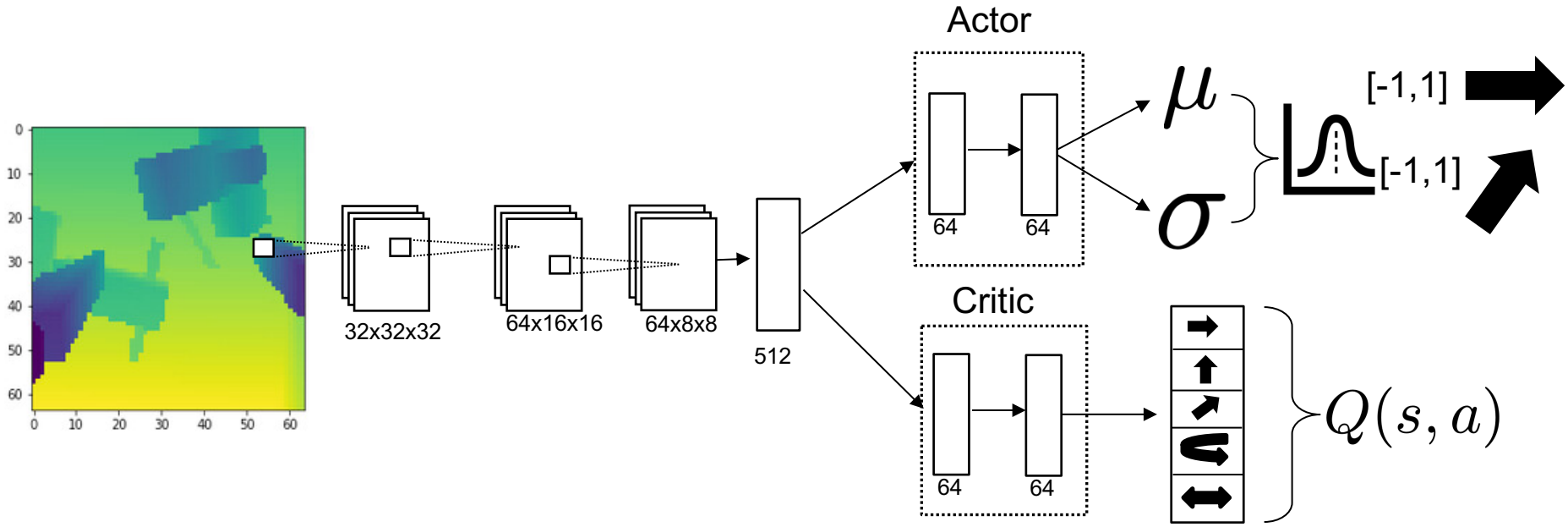



BDQ in simplified env. with increasing action dimension
Success Rate
0.2
0.4
0.6
0.8
0.0M
0.5M
1.0M
1.5M
2.0M
2.5M
3.0M
3.5M
4.0M
Number of Timesteps
BDQ_4pads
BDQ_8pads
BDQ_16pads
BDQ_33pads_small


DQN vs BDQ in simplified env.
Success Rate
0.0
0.2
0.4
0.6
0.8
0.0M
0.5M
1.0M
1.5M
2.0M
2.5M
3.0M
3.5M
4.0M
Number of Timesteps
BDQ_8pads
BDQ_33pads_small
DQN_8pads
DQN_33pads

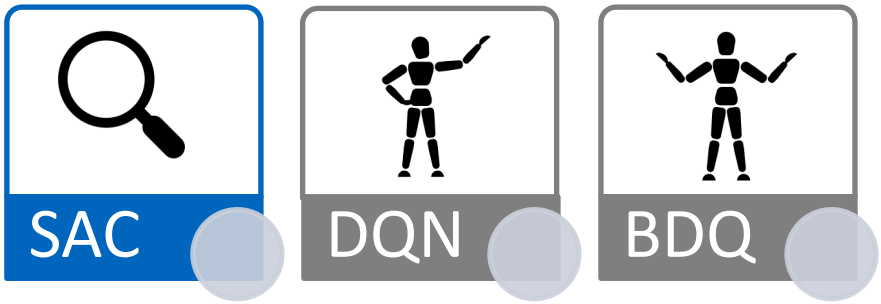


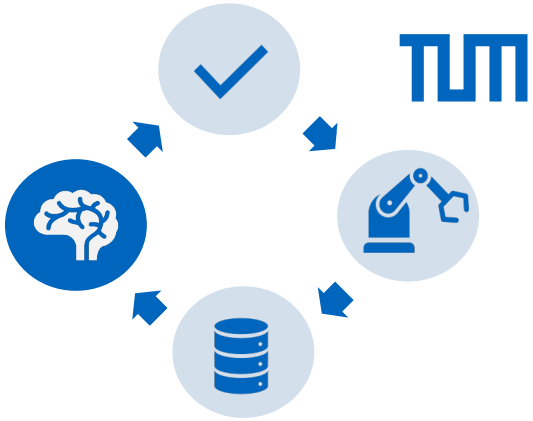


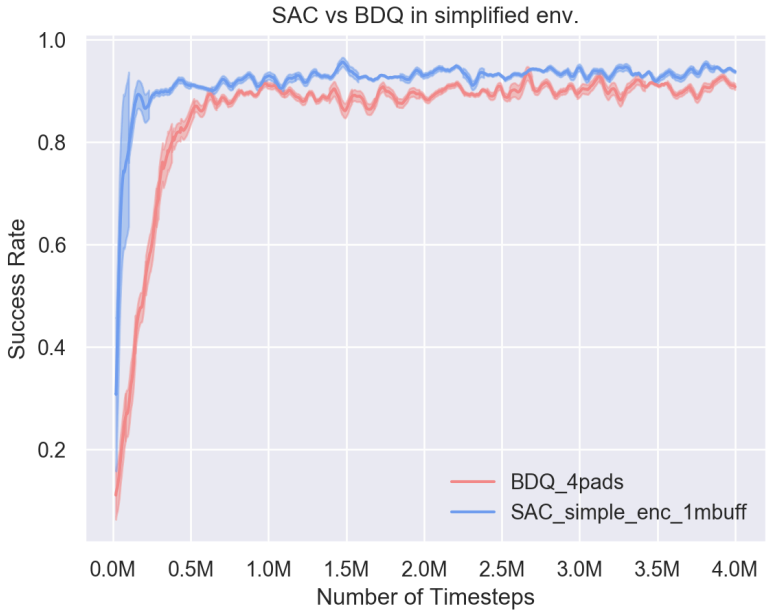


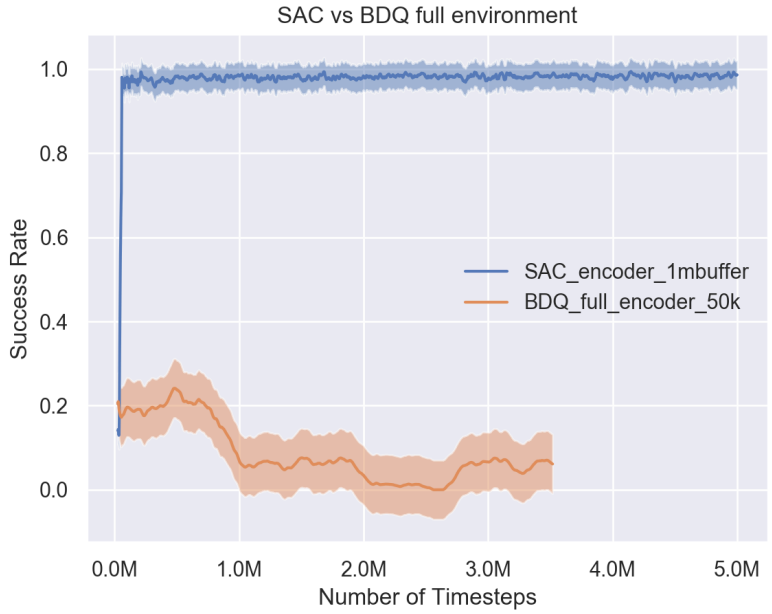
$$J(\pi) = \sum^{T} E_{(s_t,a_t)}[r(s_t,a_t) + \alpha H(\pi(.|s_t)]$$

$$V^{\pi}(s) =^{t} E_{a\sim\pi}[Q^{\pi}(s,a) - \alpha log\pi(a|s)]$$


SAC
DQN
BDQ

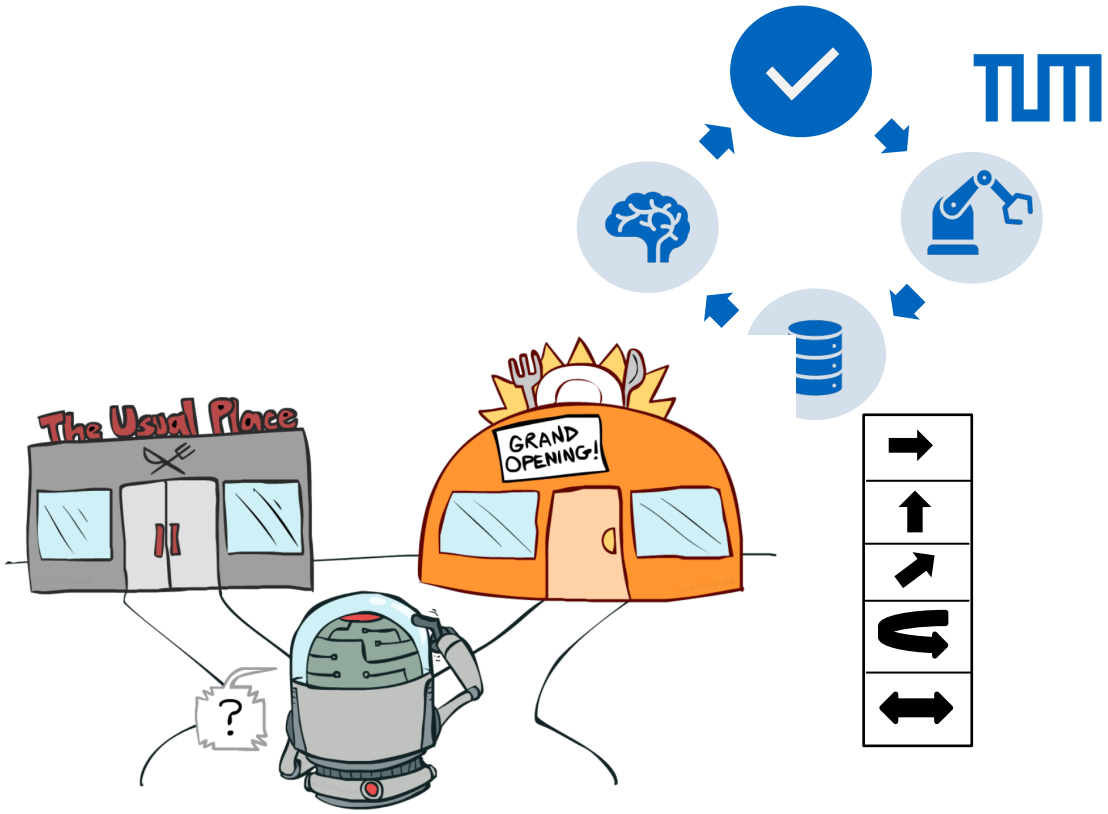



SAC vs BDQ in simplified env.
Success Rate
1.0
0.8
0.6
0.4
0.2
BDQ_4pads
SAC_simple_enc_1mbuff
0.0M
0.5M
1.0M
1.5M
2.0M
2.5M
3.0M
3.5M
4.0M
Number of Timesteps

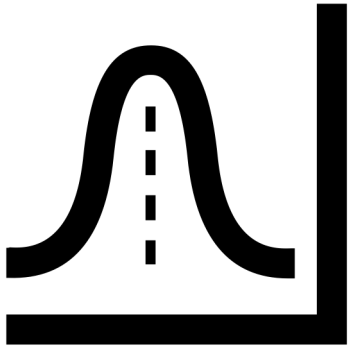

SAC vs BDQ full environment
Success Rate
1.0
0.8
0.6
0.4
0.2
0.0
SAC_encoder_1mbuffer
BDQ_full_encoder_50k
0.0M
1.0M
2.0M
3.0M
4.0M
5.0M
Number of Timesteps

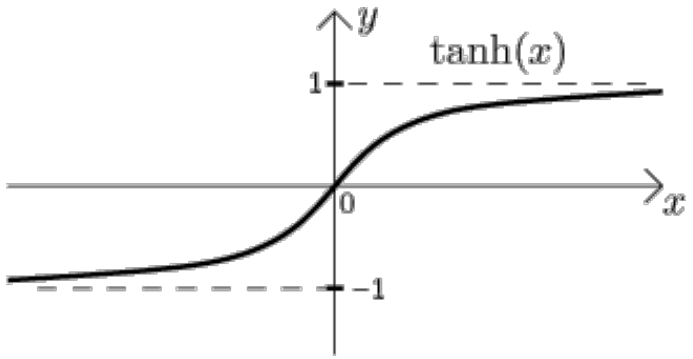
Action Selection

$$(1-\epsilon) \rightarrow argmaxQ(s,a)$$
$$\epsilon \rightarrow random\ action$$

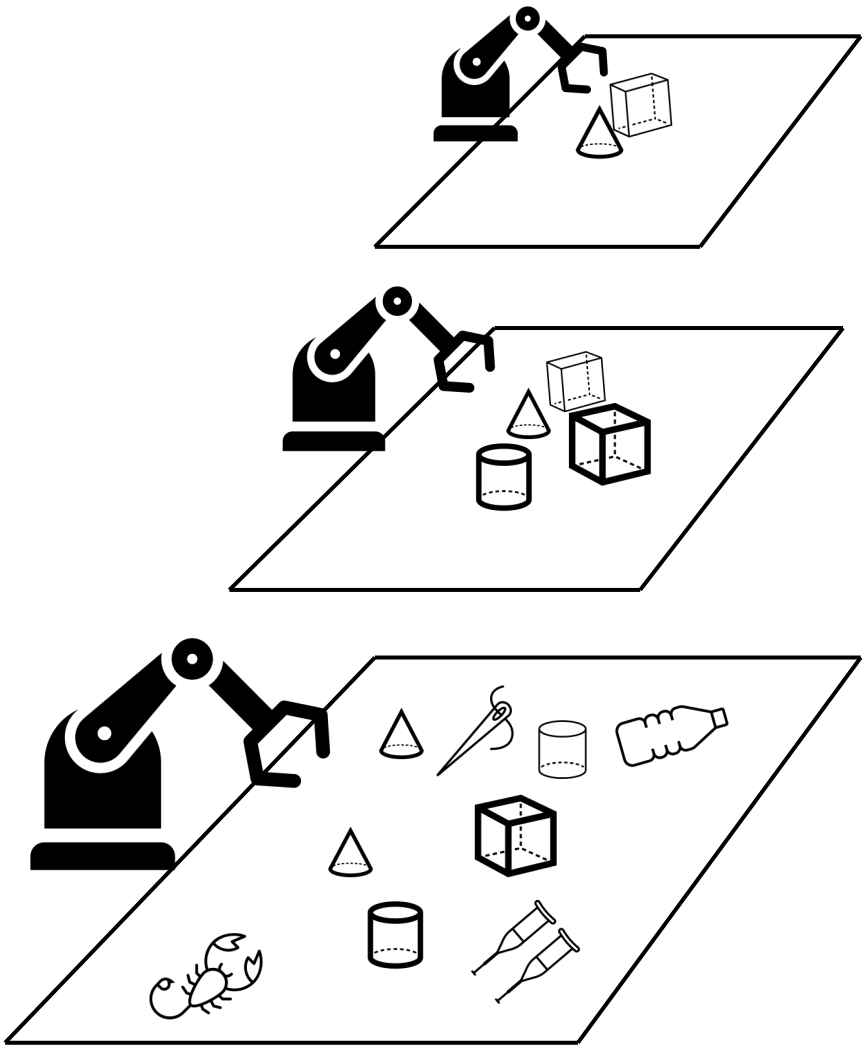


$$a_\theta(s) = \mu_\theta(s) + \sigma_\theta(s)$$

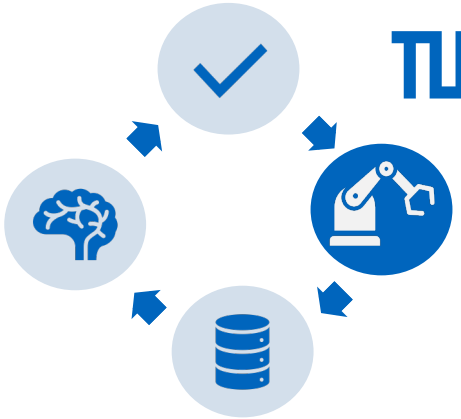


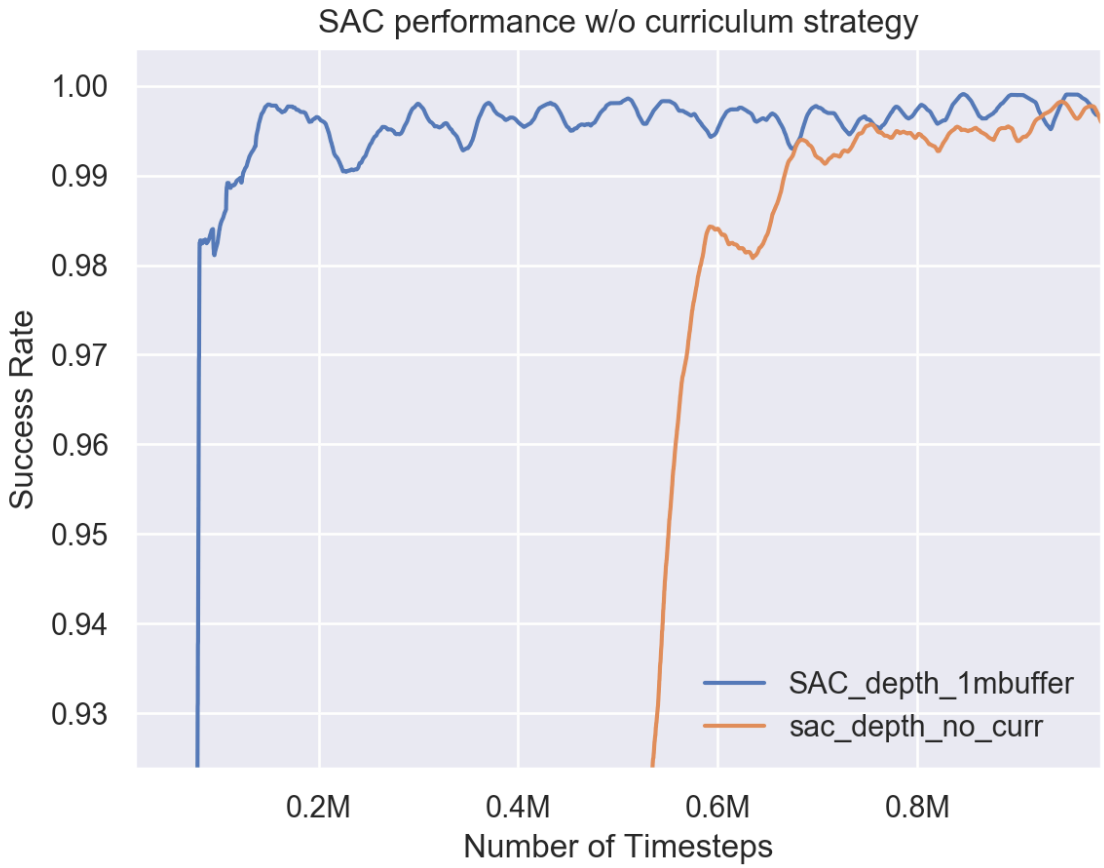


Curriculum Strategy

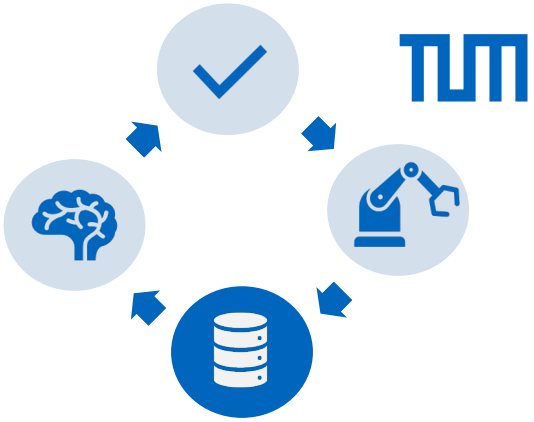


Curriculum Strategy

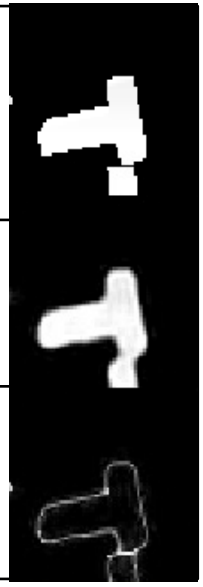




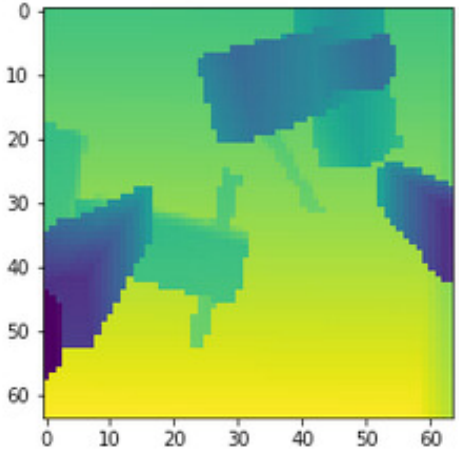
Observation



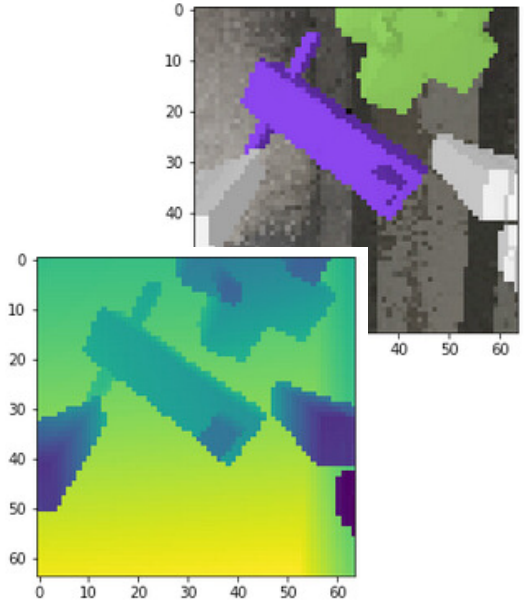
Auto-encoder

Real Image	
Recon-structed	
Error	

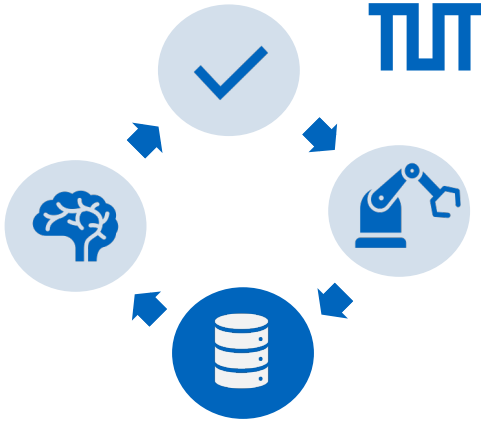
Depth

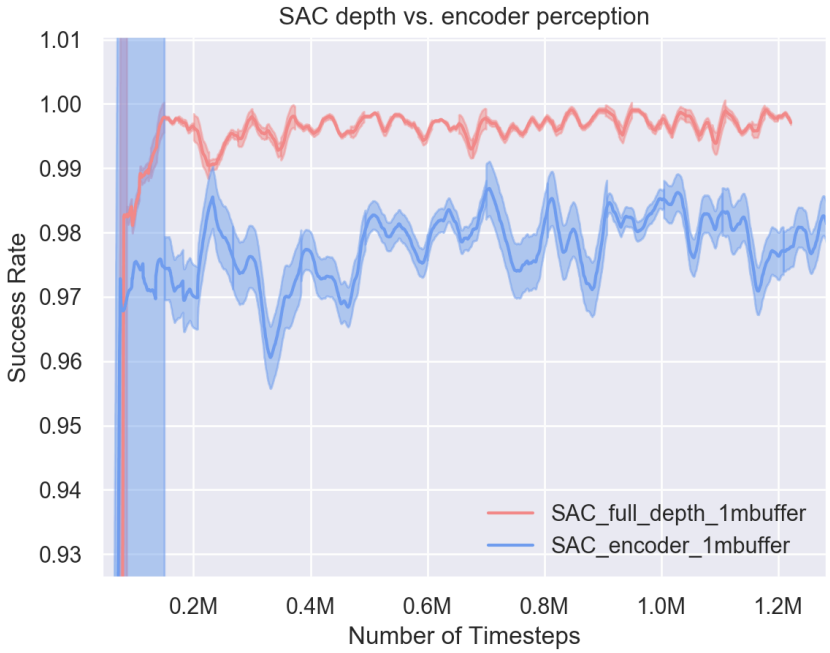


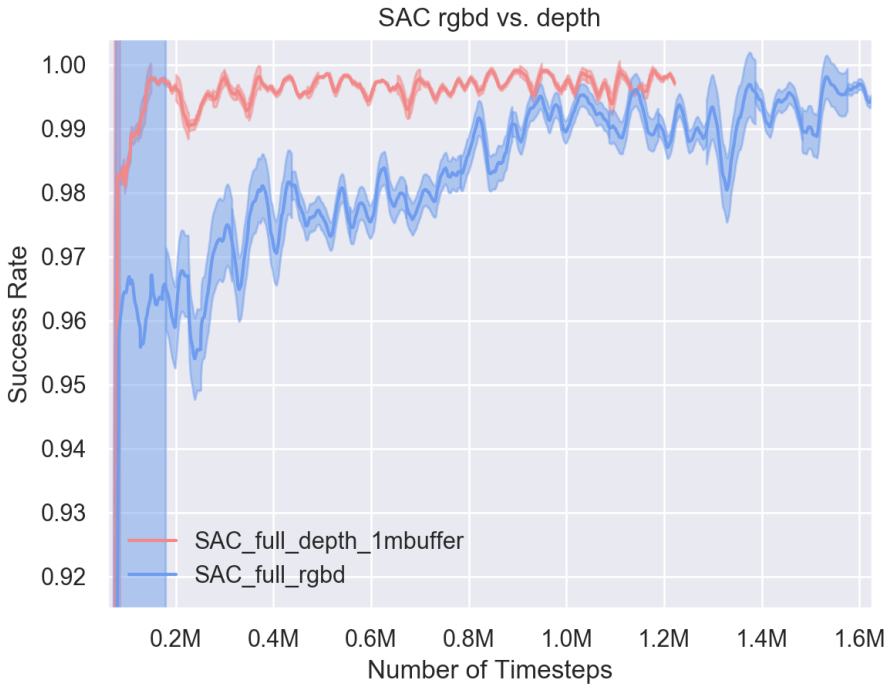
RGB-D



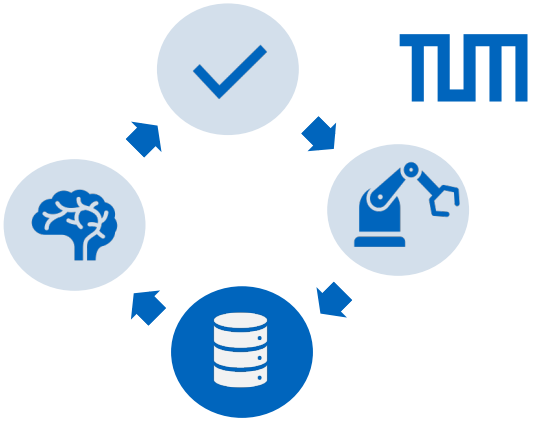
Observation

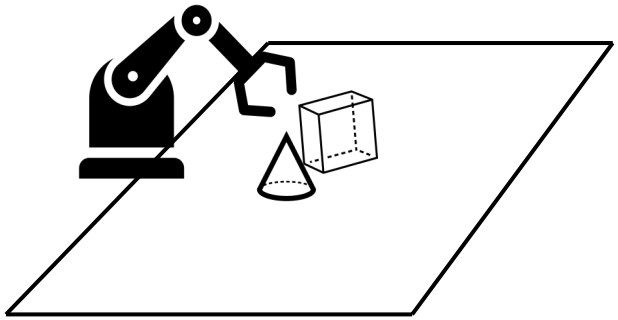




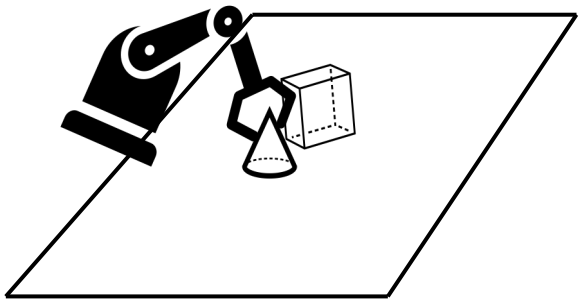


Shaped Reward

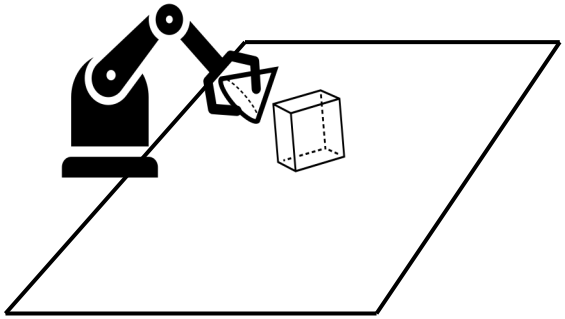




€€

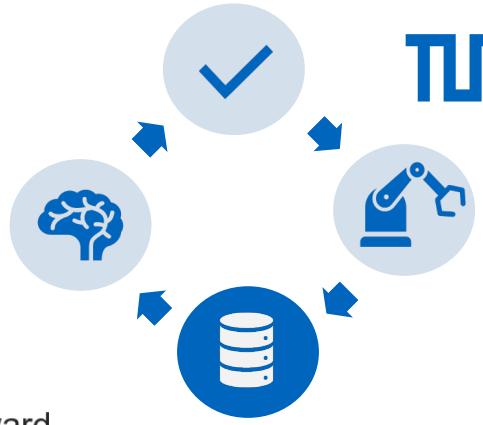


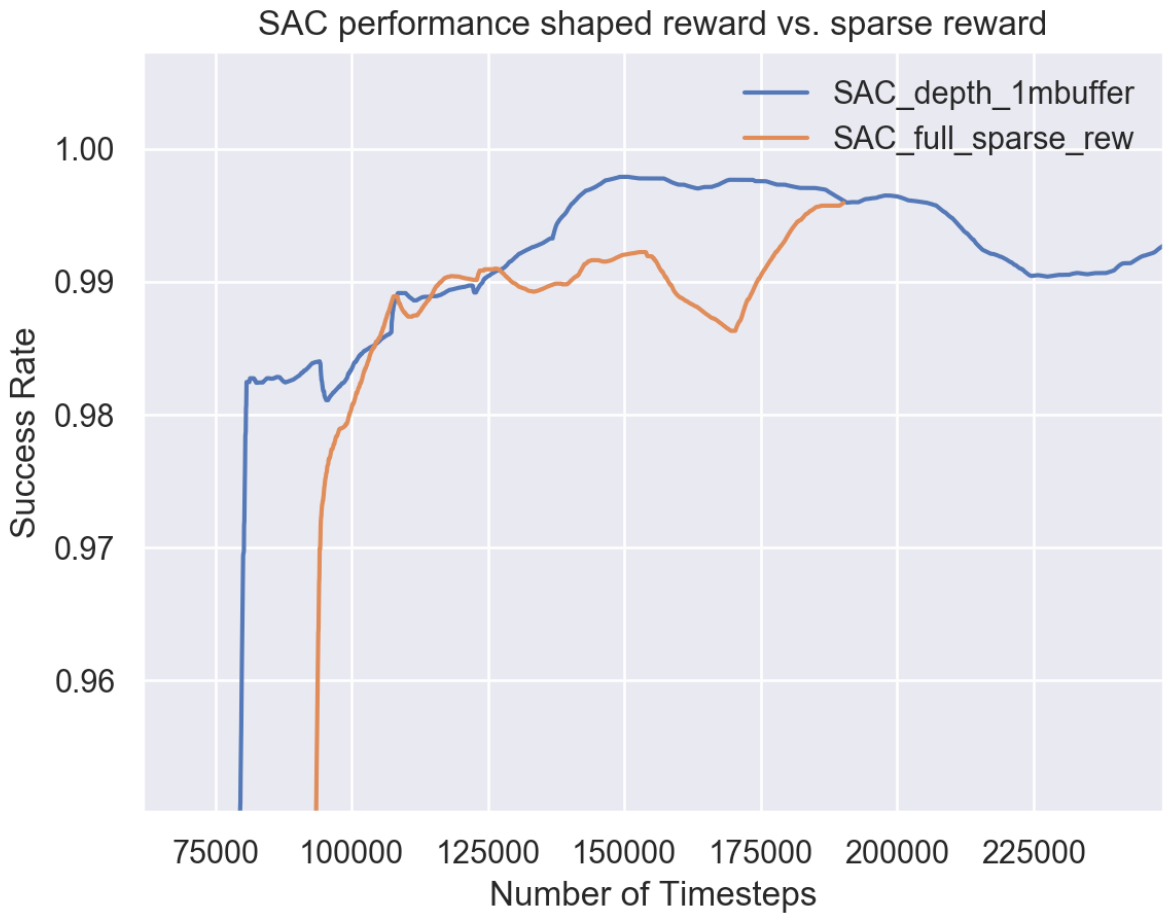
€



€€€

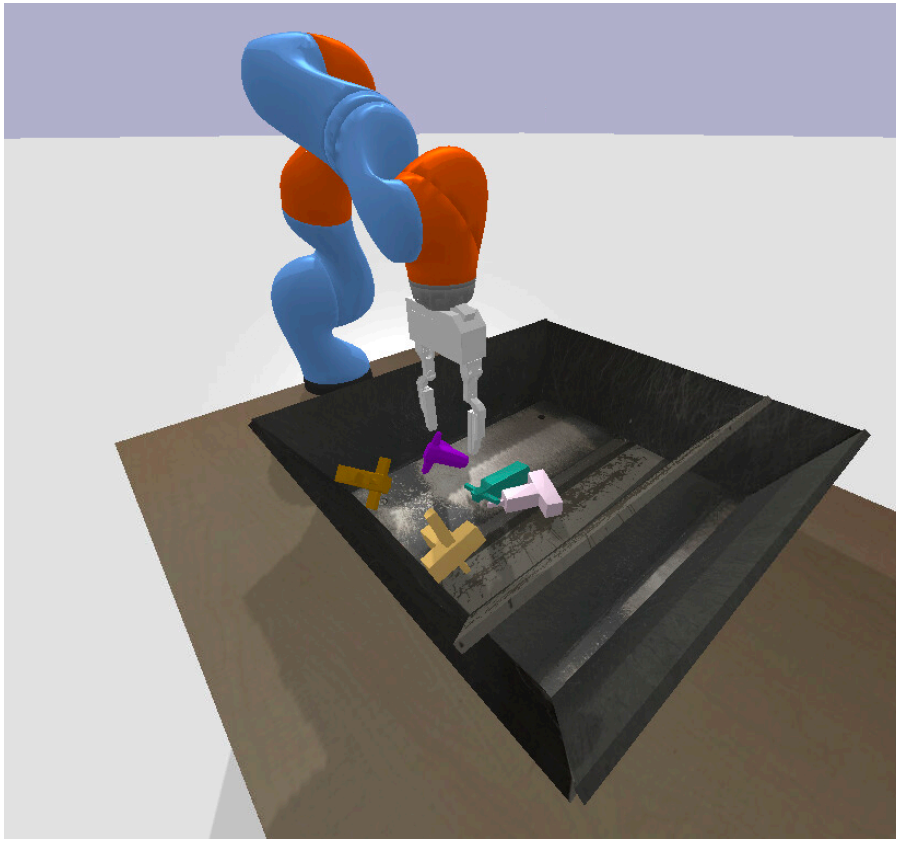
Shaped Reward





Contribution

- **Novel** grasping model
- Comparison of SAC, BDQ and DQN algorithms
- Comparison of **perception** modules



Future Timeline

- Hardware Experiments
- Multi-agent
- Targeted picking
- Soft Object Extensions

References

[1] Breyer, M., Furrer, F., Novkovic, T., Siegwart, R., & Nieto, J. (2018). Comparing Task Simplifications to Learn Closed-Loop Object Picking Using Deep Reinforcement Learning. *IEEE Robotics and Automation Letters*, *4*(2), 1549–1556. https://doi.org/10.1109/LRA.2019.2896467

[2] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

[3] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

[4] Tavakoli, A., Pardo, F., & Kormushev, P. (2018). Action branching architectures for deep reinforcement learning. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 4131–4138.

[5] Mahler, J., Matl, M., Satish, V., Danielczuk, M., DeRose, B., McKinley, S., & Goldberg, K. (2019). Learning ambidextrous robot grasping policies. *Science Robotics*, *4*(26). https://doi.org/10.1126/scirobotics.aau4984