Meta-Learning Regrasping Strategies for Physical-Agnostic Objects

Ruijie Chen[1*] Ning Gao[1,2*] Hanna Ziesche[1] Ngo Anh Vien[1] Gerhard Neumann[2]

[1]Bosch Center for Artificial Intelligence [2]Autonomous Learning Robots, KIT

Abstract

Movitation: Current visual-based grasping algorithms lack consideration of inhomogeneous physical properties of objects, e.g. different parts have different mass density or friction coefficients. Instead, predictions are conditioned purely on the geometric features.

Goal: We propose a vision-based meta-learning algorithm to learn physical properties in an agnostic way.

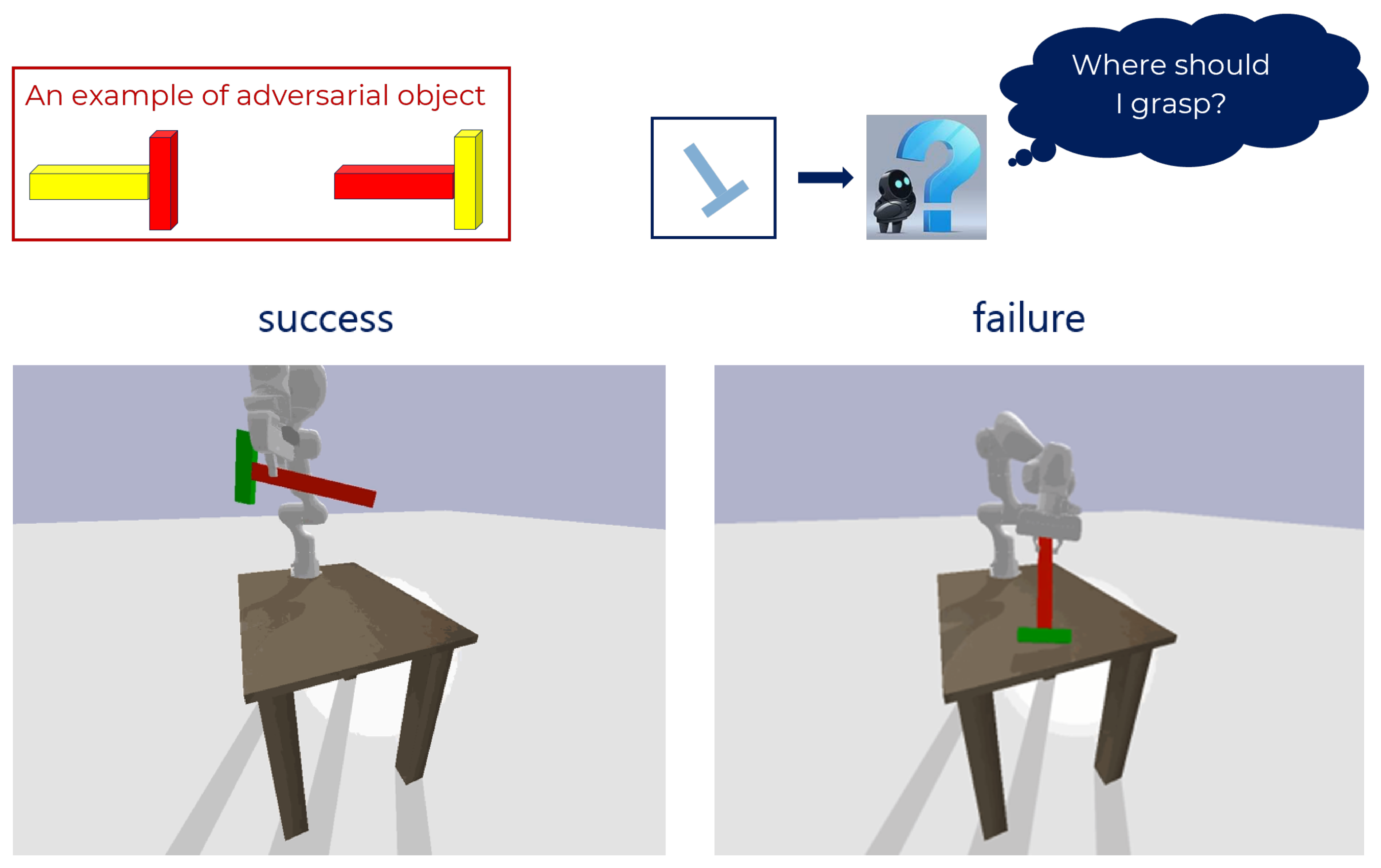
Contributions:

- Generate a categorical grasping dataset with inhomogenous mass distribution and friction coefficient of different objects.
- Propose a context-aware grasping algorithm using Conditional Neural Processes [Garnelo et al. ICML 2018].
- Evaluation on both intra- and cross-categorical unseen objects.

Problem Setting

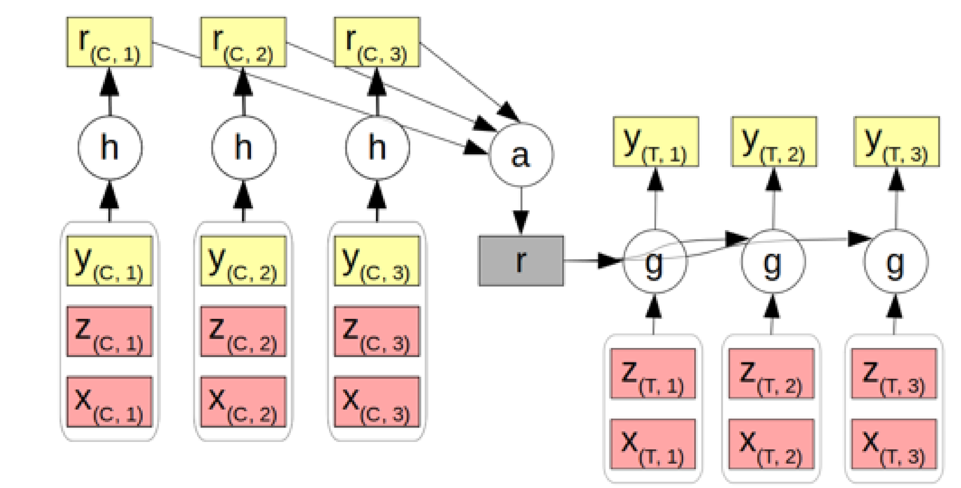
We create adversarial objects with similar shapes but different physical properties, where predictions purely depend on shapes cannot work well.

In the exemplary adversarial objects below, there are two hammer-shaped objects that are geometrically identical. For the object on the left, the hammer handle weighs 0.1kg and the hammer head weighs 0.9kg, while the hammer handle of the right object weighs 0.9 kg and the hammer head weighs 0.1kg. The visual-based algorithm such as DexNet-2.0 [Mahler et al. RSS 2017] fails at grasping because the perceived visual information of both objects looks the same.



Therefore, learning physical properties from context is essential. To achieve this, we employ Conditional Neural Processes (CNPs) to represent physical properties of each individual object implicitly from prior trials.

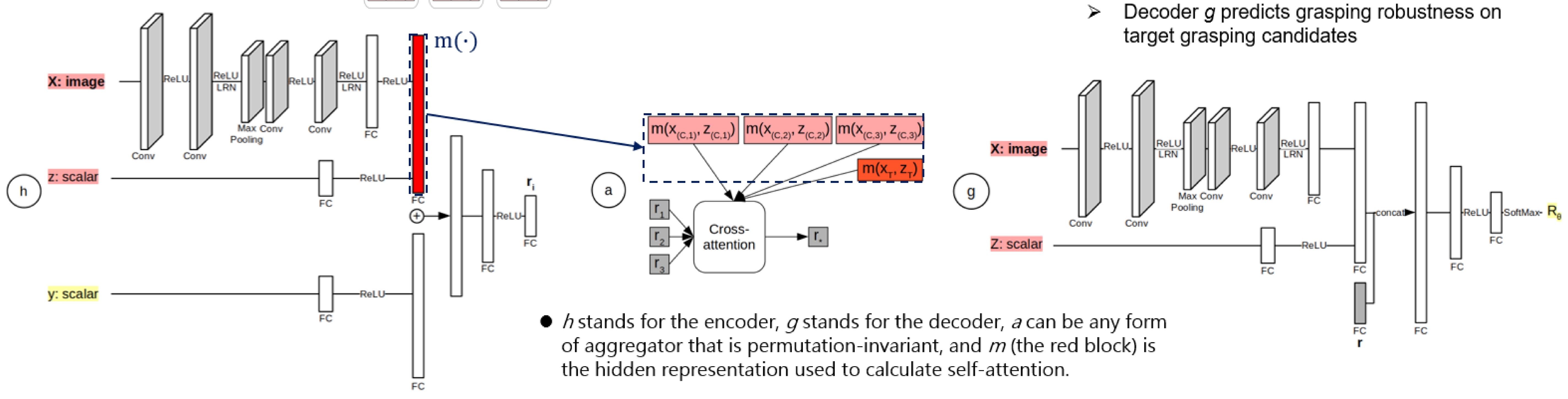
Context-Aware Grasping



$$r = \bigoplus_{i=1}^{K} h_\theta(\mathbf{x}^*_{C,i}, z^*_{C,i}, y^*_{C,i})$$

$$\hat{y}^*_T = g_\theta((\mathbf{x}^*_T, z^*_T), r) \quad y_T \quad \boxed{BCE(\hat{y}^*_T, y_T)}$$

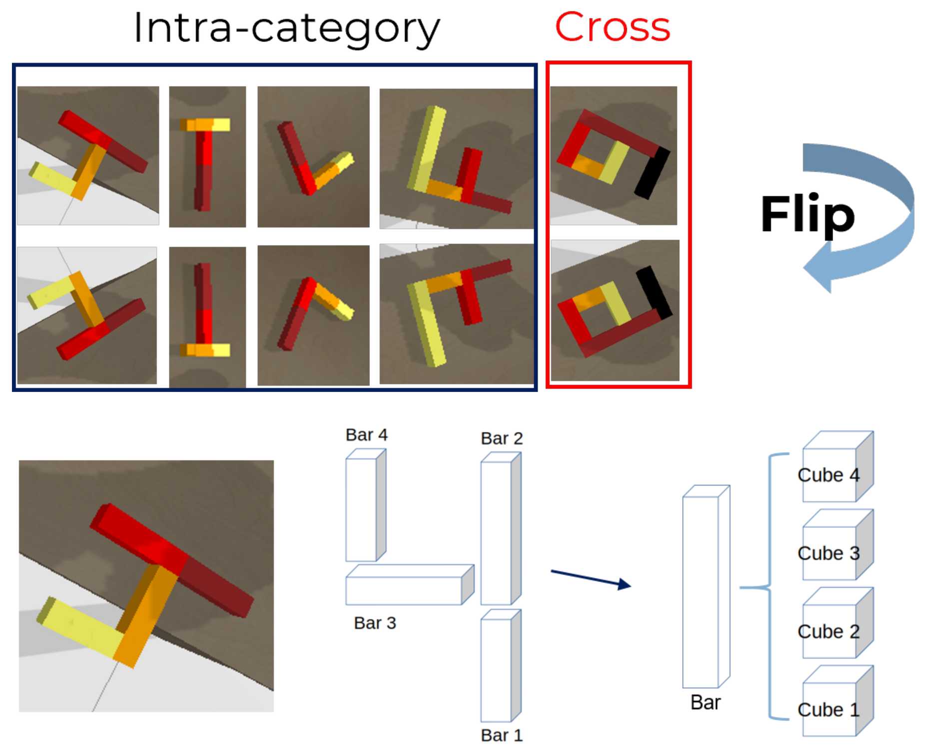
- Variables **x**, *z* and *y* are respectively
  - I. cropped depth image
  - II. grasping height
  - III. binary grasping result
- Encoder *h* extracts object representation from context observations
- Aggregator *a* gathers the extracted context information
- Decoder *g* predicts grasping robustness on target grasping candidates



- *h* stands for the encoder, *g* stands for the decoder, *a* can be any form of aggregator that is permutation-invariant, and *m* (the red block) is the hidden representation used to calculate self-attention.

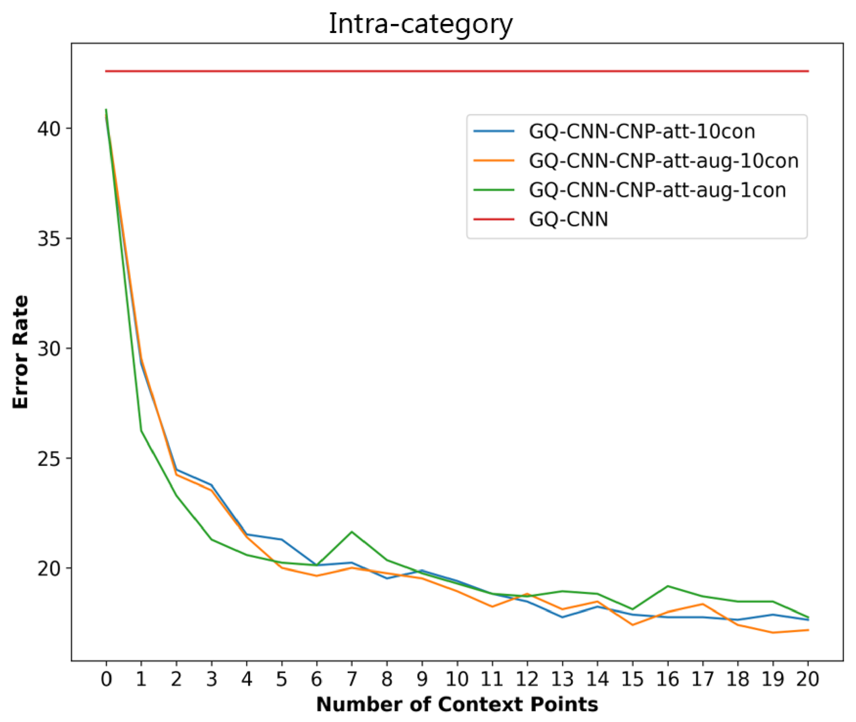
Dataset

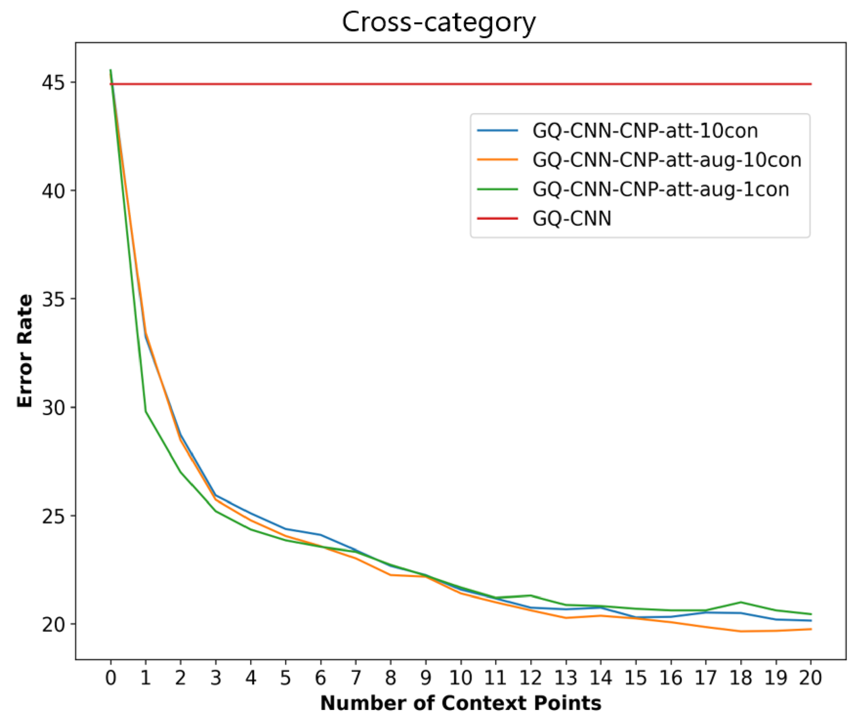
- Dataset includes 10 object categories, 5 shapes and 5 flipped shapes
- Each category includes 200-250 different objects with changing size/friction/mass of cubes
- 8 categories are used for training and intra-category evaluation and 2 unseen categories for cross-category evaluation.

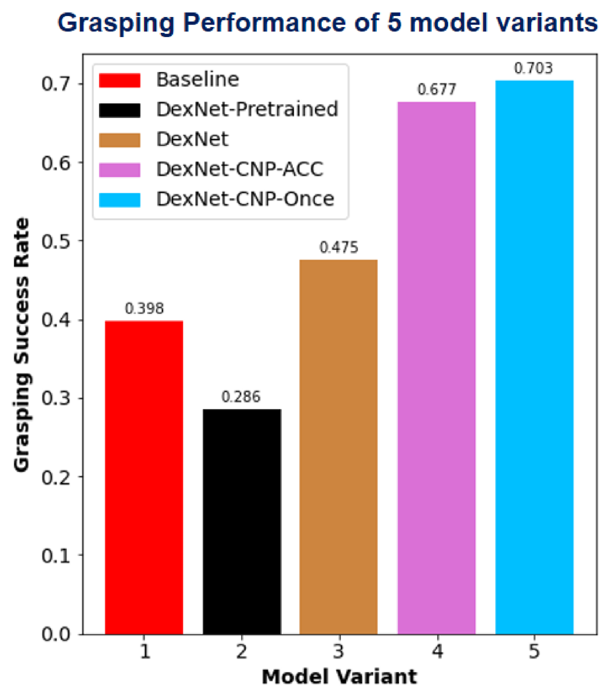


Object Variant	Total Weight	Cube Size
Letter F	1kg	(0.027, 0.033)
Letter L	1kg	(0.027, 0.033)
Letter T	1kg	(0.027, 0.033)
Number Four	1kg	(0.027, 0.033)
Number Nine	1kg	(0.027, 0.033)

Experiments & Results

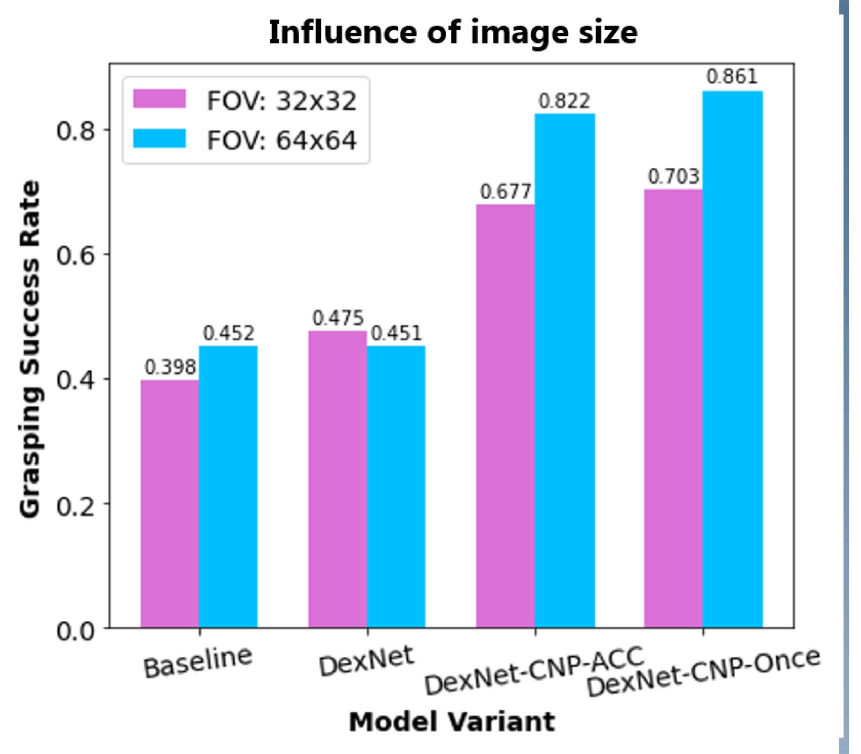


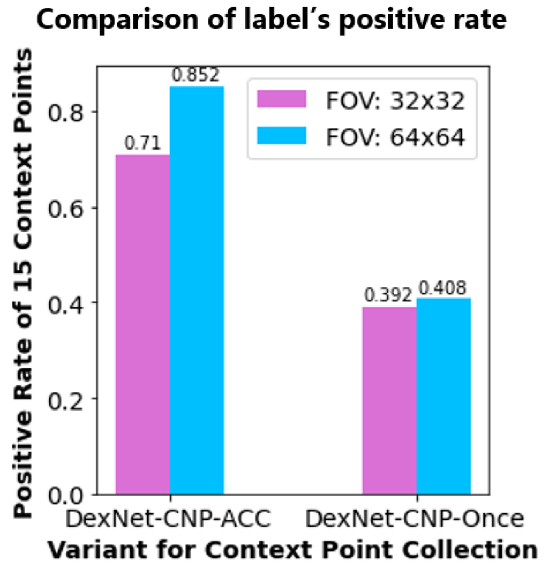


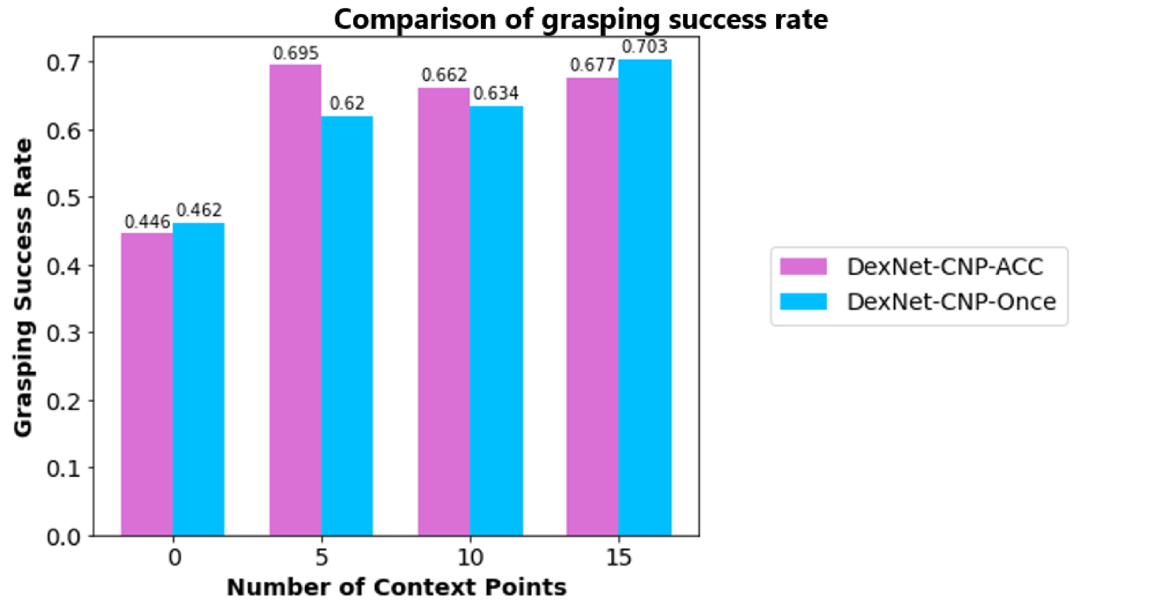


- Baseline: randomly grasp only with physical feasibility checker
- DexNet-Pretrained: DexNet with GQ-CNN trained pretrained on original dataset
- DexNet: DexNet with GQ-CNN trained on our customized dataset
- DexNet-CNP-ACC: DexNet-CNP with GQ-CNN-CNP trained on our customized dataset, where context data is collected in an accumulative way based on prior trials
- DexNet-CNP-Once: DexNet-CNP with GQ-CNN-CNP trained on our customized dataset, where context data is collected randomly

- DexNet-CNP-ACC: the collection of the next context point depends on previous ones:
  $$P(\mathbf{x}_{C,t+1}, z_{C,t+1} | (\mathbf{x}_{C,1}, z_{C,1}, y_{C,1}), \ldots, (\mathbf{x}_{C,t}, z_{C,t}, y_{C,t}))$$
- DexNet-CNP-Once: the next context point is randomly collected







Comparison of context collection strategies: Acc vs Once.

Key takeaways:

- GQ-CNN fails.
- For GQ-CNN-CNP, more context points result in more precise prediction.
- Reasonable results can be achieved given 5 context points.
- DexNet-Pretraineds performance is worse than Baseline.
- DexNet-CNP-ACC and DexNet-CNP-Once outperform DexNet by at least 20%.
- Context data collected by DexNet-CNP-ACC has much higher positive rate.
- Higher positive rate in collected context points does not necessarily help improve the grasping success rate.
- Larger FOV helps grasping performance.

More details can be found in: