A new benchmark for group distribution shifts in hand grasp regression for object manipulation. Can meta-learning raise the bar?

Théo Morales and Gerard Lacey

d-real & Science Foundation Ireland, Trinity College Dublin

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation and problem statement

Hand grasp regression methods are not commonly benchmarked for in-the-wild data or a wide variety of object grasps. They aim to generalize to unseen poses on the same objects by learning from a large collection of diverse poses and grasps. However, such models have limited use if they are only accurate for a limited set of objects. Our objective is to improve accuracy on unknown objects and grasps by:

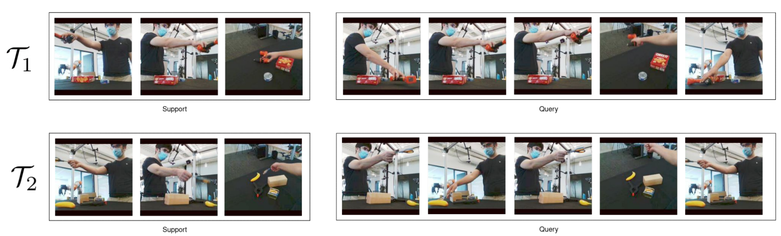
- Proposing a new benchmark for object group distribution shifts in hand and object pose regression from images.
- Reformulating grasp prediction in the context of multi-task learning such that meta-learning can be applied.

We then show that meta-learning is effective at tackling group distribution shifts for hand grasp regression. We further investigate the results with an empirical analysis to determine the directions of future work.

Benchmark creation

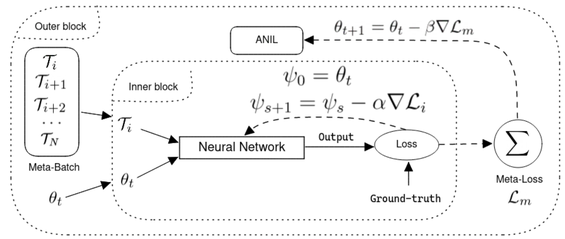
We exploit the diversity and richness of objects and grasps of the DexYCB dataset to create 9 levels of group distribution shifts. These go from easy generalisation with large diversity in the training set and poor in the test set, to hard generalisation with the opposite imbalance. The group distribution shifts are introduced by sampling non-overlapping manipulated objects for the training and test sets.

We then create a *task-based version* of this benchmark in order to train a meta-learning baseline. Each task corresponds to a support and a query set of images from one manipulation sequence of a random object from the set.



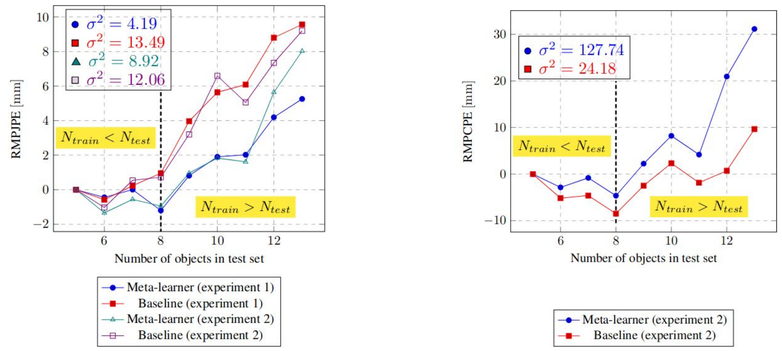
Example of tasks built from the DexYCB dataset.

Evaluating the effectiveness of meta-learning



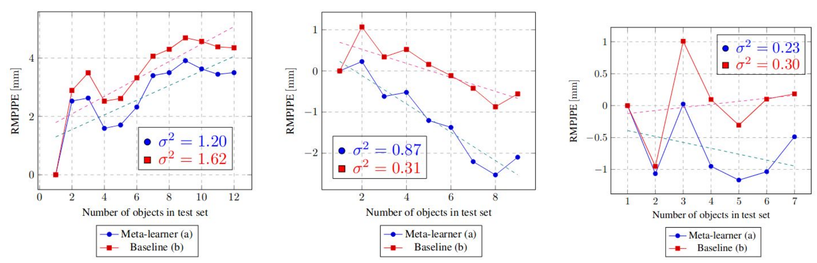
We compose **two experiments** and obtain *different results*:

1. Hand only pose prediction.
   a. The meta-learner shows significant improvements in generalisation.
   b. With enough training data, the meta-learner is more accurate on most unknown objects than the baseline.
2. Joint hand-object pose prediction.
   a. The meta-learner does not generalize significantly more for hand pose prediction.
   b. The baseline generalizes better for object pose prediction.



(a) Hand only **(b)** Object only

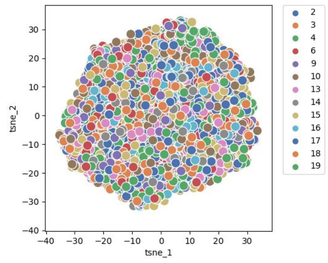
Relative Mean Per-Joint Pose Error (RMPJPE) & Mean Per-Corner Pose Error (RMPCPE) as functions of the imbalance level.



(a) 3 training objects **(b)** 6 training objects **(c)** 9 training objects

Relative Mean Per-Joint Pose Error (RMPJPE) as functions of the test split for experiment 1 with 3 training set sizes.

Empirical analysis of the results



❖ **Hypothesis 1:** the model learns specialized object-specific parameters.

Step	Tomato soup can	Banana	Scissors	Foam brick	Mug
1	204/**394**	231/**638**	270/**677**	195/**380**	236/**505**
2	193/**360**	219/**576**	255/**613**	185/**338**	225/**458**
3	182/**330**	207/**522**	242/**557**	176/**302**	214/**416**
4	173/**303**	197/**475**	229/**509**	167/**273**	205/**381**
5	164/**280**	187/**435**	217/**467**	159/**248**	195/**349**
6	155/**259**	178/**401**	207/**430**	152/**227**	187/**322**
7	148/**241**	169/**371**	197/**398**	145/**209**	179/**298**
8	141/**225**	161/**344**	188/**369**	139/**194**	172/**277**
9	134/**210**	153/**321**	180/**344**	133/**181**	165/**259**
10	128/**198**	146/**301**	172/**322**	127/**169**	159/**243**

❖ **Hypothesis 2:** Using Oriented Bounding Box coordinates in the training signal constrains the hand-object pose.

Conclusions

We observe that meta-learning is:

- Better at dealing with group distribution shifts for hand pose prediction.
- More accurate on most unknown objects than the baseline for hand pose prediction.
- Ineffective for joint hand-object pose regression.

In future work we aim to:

- Explicitly formulate the objective to encourage specialization and reduce meta-overfitting.
- Express the hand-object constraints in the objective.
- Apply this method to the SOTA.

References

Chao et al. (2021). "DexYCB: A Benchmark for Capturing Hand Grasping of Objects." In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR): 9040-9049

Raghu et al. (2020). "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML." In: International Conference on Learning Representations.

Huang et al. (2021) "Survey on depth and RGB image-based 3D hand shape and pose estimation." In: Virtual Real. Intell. Hardw. 3, pp. 207-234.