Supplementary for M2T2: Multi-Task Masked Transformer for Object-centric Pick and Place

Anonymous Author(s)
Affiliation
Address
email

1 Further Details on M2T2 archiecture

The contact decoder in M2T2 has three sets of learnable embeddings: two task embeddings, G grasp embeddings, P placement embeddings. The $G + P$ query tokens are added with learnable position encodings and fed into a transformer network with nine blocks. Each block consists of a cross-attention layer, a self-attention layer and a feedforward MLP layer. In the cross-attention layer, the query tokens are cross attended with one of the feature maps produced by the scene encoder to incorporate scene context. The three context feature maps are passed in turn to the nine transformer blocks, so that each feature map is cross-attended three times. The input tokens of each transformer block are also used to produce intermediate predictions, which are used as attention masks for the cross-attention layer of the next block. This forces the network to focus on relevant regions in the scene.

While the M2T2 design is inspired by [1], we adopt it to handle 3D inputs (instead of images) and multi-task outputs. For example, since the context features are grounded to 3D points, we add position encodings computed from their 3D locations to the context features before feeding them into cross-attention.

2 Examples of Synthetic Data

We procedurally generated a large-scale synthetic dataset for training M2T2, as shown in Fig 1.

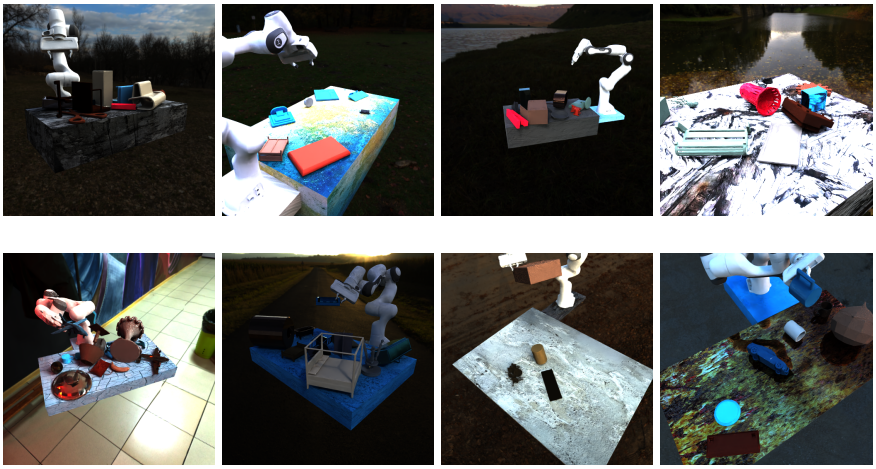

Figure 1: Examples for our large-scale synthetic dataset, for the grasping (top) and placing (bottom) tasks respectively. Note that the textures and color images are just for display. Only the corresponding depth images were used for training. Objects are randomly sampled from ACRONYM [2] and placed in their stable poses.

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

References

[1] B. Cheng, A. Choudhuri, I. Misra, A. Kirillov, R. Girdhar, and A. G. Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[2] C. Eppner, A. Mousavian, and D. Fox. ACRONYM: A large-scale grasp dataset based on simulation. In *2021 IEEE Int. Conf. on Robotics and Automation, ICRA*, 2020.