WHERE ARE MY BOXES?

SELF-SUPERVISED DETECTION AND POSE ESTIMATION OF LOGISTICAL OBJECTS IN 3D SENSOR DATA

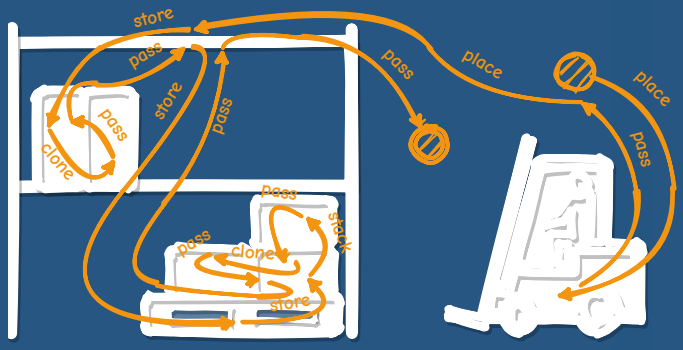
ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Nikolas Müller,
Jonas Stenzel,
Jian-Jia Chen,

German Research Center for Artificial Intelligence
Fraunhofer Institute for Materialflow and Logistics
Technical University Dortmund

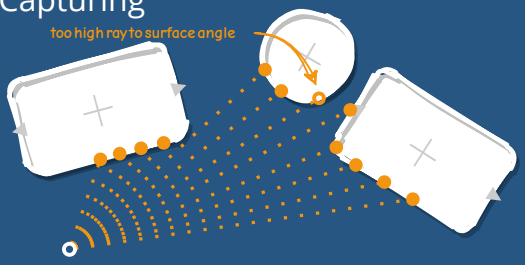
For full process automation in Industry 4.0 with smart machines, localizing objects is a crucial capability. Enabling such an object localization with Machine Learning is tempting, as it avoids a manual engineering of features for every object class, yet gathering large amounts of accurate pose data to train neural networks can be just as difficult.

In this work, a novel **self-supervised approach** based on point clouds is presented, which resolves these issues and allows a robust detection and localization of objects in cluttered scenes, while compensating for noise, occlusions and symmetries. It leverages **3D sensor data simulation** being simpler than light rendering to construct random scenes using a stochastic process by maintaining object relations with placement operations like cloning, stacking and storing, and to then train a **fully convolutional voting network** with random scans from those scenes.



In logistical scenes, objects of the same class tend to be placed next to each other, while stackable objects tend to be stacked, and storable objects tend to be placed inside a valid storage. To maintain these relations, the stochastic construction process itratively builds a scene by cloning, stacking, and storing objects with reference to an already placed object using multivariate normal distributions.

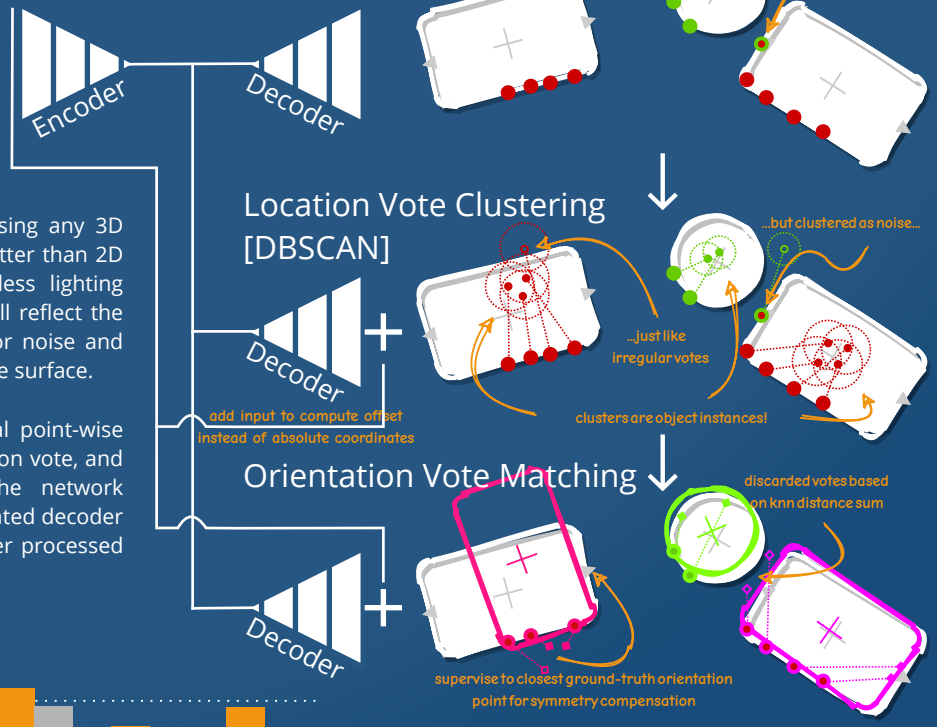
Capturing



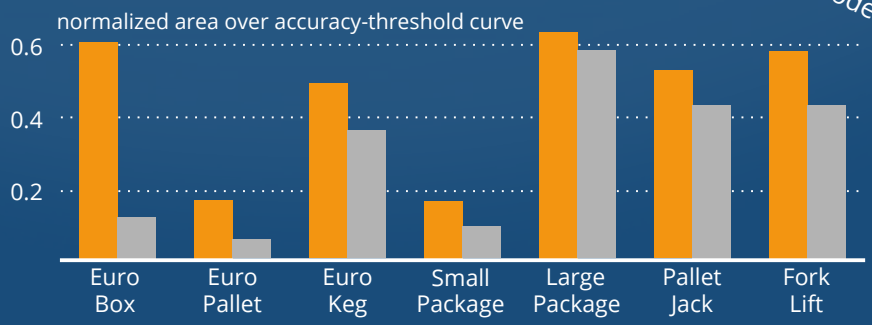
The randomly generated scenes can be captured using any 3D sensor. Even though 3D data is generalizing much better than 2D data due to more accurate receptive fields and less lighting disturbances, the simulated sensor model should still reflect the real sensor as close as possible. This includes sensor noise and effects like missing values on too high ray angles on the surface.

The captures are then fed into a fully convolutional point-wise voting network to predict a classification, object location vote, and orientation point vote for each input point. The network architecture is based on KPFCNN [1], but with a replicated decoder for each of the three outputs. The outputs are further processed into object poses as shown on the right.

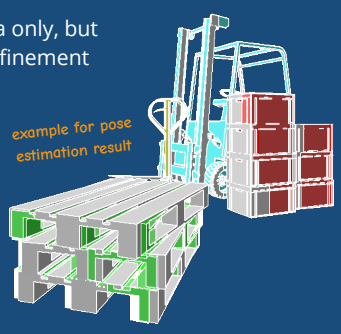
Class Partitioning



Results



ours
PVN3D [2] on 3D data only, but with iterative pose refinement



[1] Thomas, Hugues, et al. "Kpconv: Flexible and deformable convolution for point clouds." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[2] He, Yisheng, et al. "PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH
Fraunhofer IML
technische universität dortmund