

CroCPS - Supplementary Material

BMVC 2022 Submission # 390

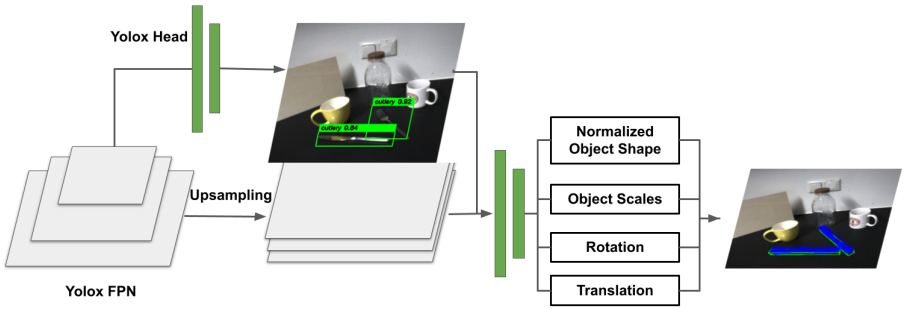


Figure 1: The 3D lifting module to derive the object 6D poses and sizes. The YoloX [1] FPN features are upsampled to the same scale and fed into the network to estimate the normalized object shape, scales, translation and rotation of the objects.

1 Implementation Details

The 3D lifting module is visualized in Fig. 1. The YoloX [1] Nano model is leveraged for the 2D detections, where the FPN features of sizes $64 \times 60 \times 80$, $128 \times 30 \times 40$, $256 \times 15 \times 20$ are upsampled and concatenated as features of size $448 \times 60 \times 80$. The ROI features inside the 2D bounding boxes are extracted by ROI alignment and used for 3D lifting. For the self-supervision of category-level object poses and sizes in real images, the mask loss, the normal loss and the geometric loss are multiplied with factors of 50, 50, 1000. The self-supervision network is trained for 15000 iterations with a SGD optimizer and a base learning rate of $1e-5$.

2 Visualization of the Synthetic Dataset

In total 10k synthetic images are generated for the pre-training of the network. Examples are visualized in Fig. 2.

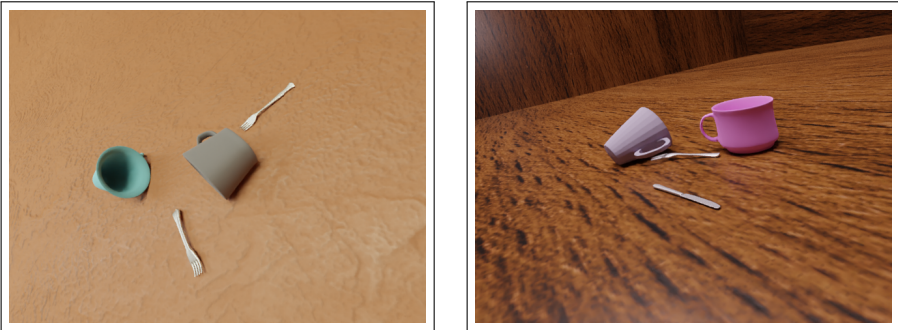


Figure 2: Visualization of rendered synthetic images

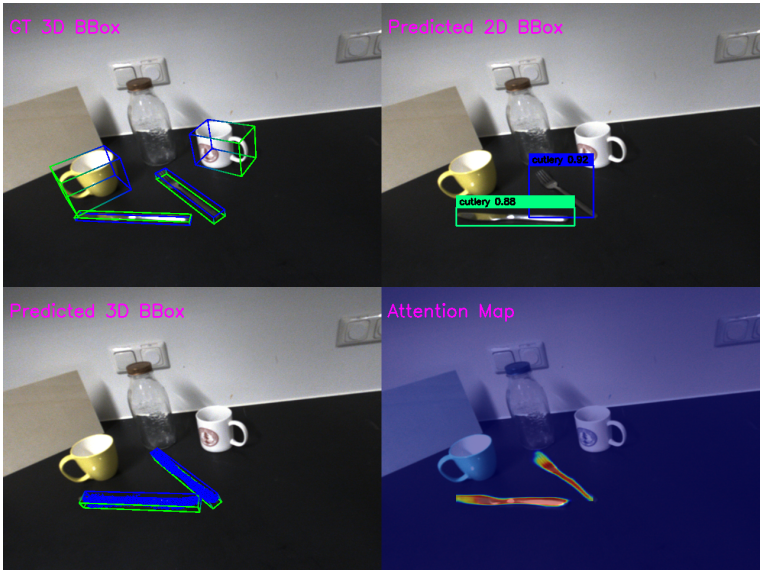


Figure 3: Visualization of detection results after self-supervision

3 Qualitative Detection Results

Qualitative results after self-supervision are illustrated in Fig. 3. The ground truth 3D bounding box of cups and cutlery, the predicted 2D bounding boxes, the predicted 3D bounding boxes of cutlery, along with attention maps, are visualized.

References

[1] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.