Object Level Depth Reconstruction for Category Level 6D Object Pose Estimation From Monocular RGB Image

Zhaoxin Fan [1], Zhenbo Song [2], Jian Xu [4], Zhicheng Wang [4], Kejian Wu[4]], Hongyan Liu[3]

[1] Renmin University of China, [2] Nanjing University of Science and Technology
[3] Tsinghua University, [4] Nreal

Arxiv Link: https://arxiv.org/pdf/2204.01586.pdf

Motivation

1. Depth information prohibits broader applications.
2. Reconstructing metric-scale mesh makes the pipeline redundant.
3. Reconstructing the object-level depth preserves shape details.

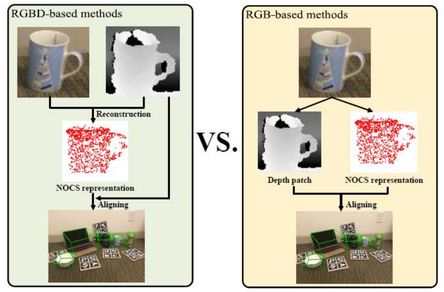


Figure 1. Difference between RGBD-based methods and our RGB-based method. RGBD- based methods take RGB image and depth channel as inputs, and the output is a canonical NOCS representation. While our RGB-based method only takes RGB images as input, and predict the NOCS representation as well as the object-level depth simultaneously..

Contribution

1. We propose OLD-Net, a novel deep learning approach for category-level 6D object pose estimation, which aims at directly predicting object-level depth from a monocular RGB image in a simple yet effective way.
2. We propose the Normalized Global Position Hints and the Shape-aware De-coupled Depth Reconstruction scheme. Both modules are tailored for RGB-based category-level 6D object pose estimation.
3. We conduct extensive experiments on two challenging datasets to verify the effectiveness of our method. Our model achieves state-of-the-art performance in both synthetic and real world scenarios.

Ours: Pipeline of our work and OLD-Net

The bottom figure is the pipeline of this paper, we first train a Detect-Net and an encoder-decoder network to crop image patches and generate the shape prior respectively. Then, we predict the object-level depth and the NOCS representation. 6D object pose is recovered by aligning the depth and the NOCS representation. To predict high-quality object-level depth, we design the novel OLD-Net as shown in the top figure.

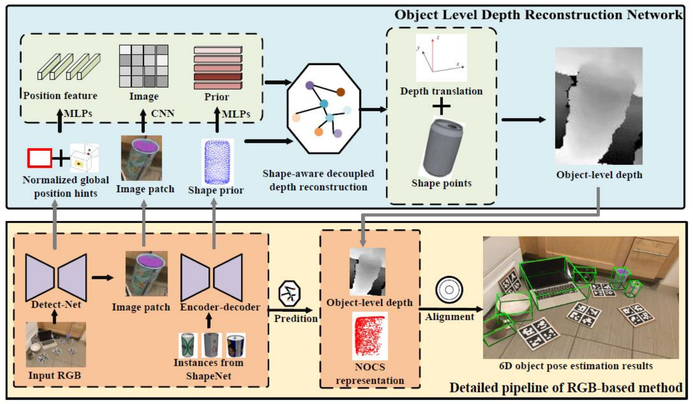
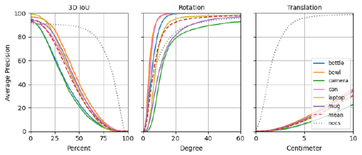
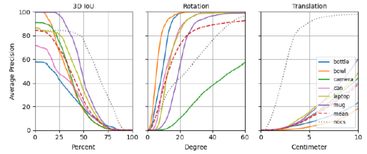


Figure 2. Pipeline of our work and OLD-Net.

Results:



(a) Results on CAMERA25 dataset



(b) Results on REAL275 dataset

Figure 3. The average precision (AP) vs. different thresholds on 3D IoU, rotation error, and translation error. our method performs excellently in terms of IoU and rotation on all categories. Our model even achieves comparable performance with the RGBD based method in terms of rotation prediction.

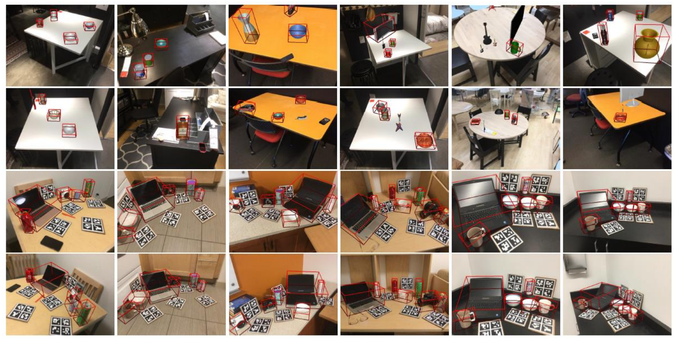


Figure 4. Qualitative results of successful cases. Top two rows are results on CAMERA25 dataset and bottom two rows are results on REAL275 dataset. Our method can accurately estimate the pose and size of the object taking a single RGB image as input..

Conclusion

In this paper, we propose a novel network named OLD-Net for RGB-based category-level 6D object pose estimation. Directly predicting object-level depth using shape prior is the key insight of our work. To reconstruct high-quality object-level depth, we introduce the Normalized Global Position Hints and Shape-aware Decoupled Depth Reconstruction Scheme in OLD-Net. We also predict the canonical NOCS representation of the object in our pipeline using adversarial training. Extensive experiments on both real and synthetic datasets have demonstrated that our method can achieve new state-of-the-art performance.

Acknowledgments

This work was supported in part by National Key Research and Development program of China under Grant No. 2020YFB2104101 and National Natural Science Foundation of China (NSFC) under Grant Nos. 62172421, 71771131, and 62072459.