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

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

[1] Key Laboratory of Data Engineering and Knowledge Engineering of MOE, School of Information, Renmin University of China, 100872, Beijing, China
{fanzhaoxin,hejun}@ruc.edu.cn
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, 210094, Nanjing, China
songzb@njust.edu.cn
[3] Department of Management Science and Engineering, Tsinghua University, 100084, Beijing, China
hyliu@tsinghua.edu.cn
[4] Nreal
{jianxu,zcwang,kejian}@nreal.ai

1 Appendix

1.1 Implementation Details

We implement our method by PyTorch and optimize it using the Adam optimizer. The image patch is resized to 192×192. We randomly select 1024 pixels to predict depth during training. The Detect-Net is Mask-RCNN [1]. A PSP-Net [6] with a ResNet-18 [2] backbone is used to learn image features. Points number of the shape prior is 1024. We set $C = 64$ and $C_g = 1024$. The model is trained for 50 epochs with a batch size of 96. The initial learning rate of the main network is 0.0001 with a decay rate of 0.1 at the 40th epoch. The initial learning rate of the discriminator is 0.00001. It also decays by 0.1 at the 40th epoch. The balance terms γ_1 to γ_7 are 1.0, 0.1, 0.1, 1.0, 5.0, 0.0001 and 0.01. We train our model on a single RTX 3090 GPU. Note we also have to recover the object size, we simply use the average size of $P_{pri} + D_{nocs}$ as our result following [4]. Following [3], we report the mean Average Precision (mAP) metric. Six kinds of mAPs are chosen. They are mAP at $IoU > 0.25$ (IoU25), mAP at $IoU > 0.5$ (IoU50), mAP at $IoU > 0.75$ (IoU75), mAP at $translation < 10cm$ (10cm), mAP at $rotation < 10°$(10 °) and mAP at the threshold of 10 °10 cm.

⋆ Corresponding author

1.2 Datasets

To verify the effectiveness of our method. We conduct experiments on the CAMERA25 dataset [5] and REAL275 dataset [5]. They are currently the most prevalent benchmark datasets for category-level 6D object pose estimation. The CAMERA25 dataset is a synthetic dataset that contains 300K RGBD images (with 25K for evaluation) generated by rendering and compositing synthetic objects into real scenes. The REAL275 dataset is a real-world dataset that contains 4.3K real-world RGBD images from 7 scenes for training, and 2.75K real-world RGBD images from 6 scenes for evaluation. Both datasets consist of six categories, i.e., bottle, bowl, camera, can, laptop and mug. Note though they provide RGBD images, we only use the RGB part of these images to predict the 6D object pose during evaluation.

1.3 More results

To help readers to better understand our work, we visualize more results in Fig. 1 and Fig. 2. Fig.1 shows some results on the CAMERA25 dataset and Fig. 2 shows some results on the REAL275 dataset. Both successful cases and failure cases are included. We hope readers can refer to these visualizations to find more observations that may benefit future works.

From the failure cases we can find that that OLD-Net may miss objects or detect ghosts sometimes. This may be solved by using a stronger Detect-Net. Researching 2D detection is out of the scope of this paper and we leave it as a future work. Another observation from failure cases is that we find sometimes our model can not excellently recover the size of the object. Therefore, further future efforts should also be made to improve the object size estimation accuracy.

We believe that our work is a significant step to enable RGB-based category-level 6D object pose estimation to be deployed into many potential applications like robotics and augmented reality. Besides solving the limitations, there are also some future works we suggest to do: 1) Semi-supervised training on both labelled synthetic data and unlabelled real world data. 2) Designing stronger object-level depth prediction network architectures. 3) Trying domain adaptation methods.

References

1. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Lee, T., Lee, B.U., Kim, M., Kweon, I.S.: Category-level metric scale object shape and pose estimation. IEEE Robotics and Automation Letters **6**(4), 8575–8582 (2021)
4. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: European Conference on Computer Vision. pp. 530–546. Springer (2020)

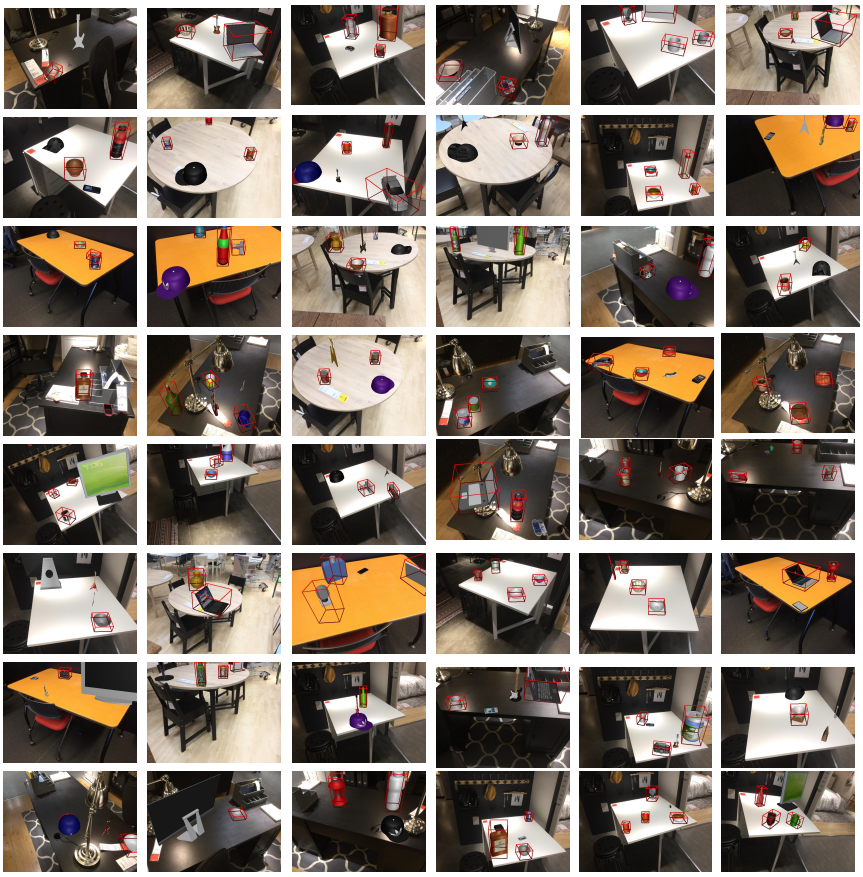

Fig. 1. More visualization results on CAMERA25 dataset.

5. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)
6. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

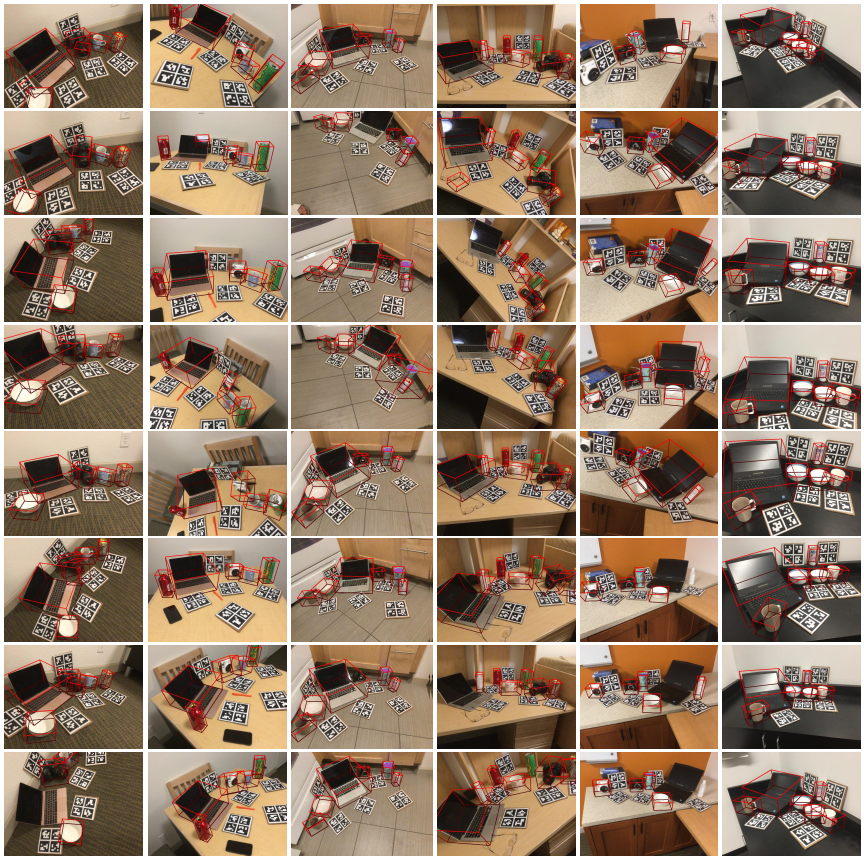

Fig. 2. More visualization results on REAL275 dataset.