Supplementary Material
DGECN: A Depth-Guided Edge Convolutional Network for End-to-End 6D Pose Estimation

Tuo Cao[1], Fei Luo[1]*, Yanping Fu[2], Wenxiao Zhang[1], Shengjie Zheng[1], and Chunxia Xiao[1]*
[1]School of Computer Science, Wuhan University, Wuhan, Hubei, China
[2] School of Computer Science and Technology, Anhui University, Hefei, Anhui, China
ypfu@ahu.edu.cn, wenxxiao.zhang@gmail.com, zsj_mdk@163.com, {maplect,luofei,cxxiao}@whu.edu.cn

Abstract

In this supplementary material, we first elaborate the details about our network architecture in Section A. Then, running time analysis is given in Section B. Finally, more visual results are given in Section C. Note that we did not include all the material in the main paper due to the space limit.

1. A. Detail about Network Architecture.

We feed DGECN with a RGB image similar to PVNet [5] and PoseCNN [6] and directly output 6D pose. After a cross-domain feature fusion block, we leverage SegPose [3] as backbone to estimate 2D-3D correspondences from the multi-fusion feature of size 256×256. Finally, DG-PnP directly estimates the 6D pose from the estimated 2D-3D correspondences. We set $\lambda_{1-4} = 1$ in formula 4 in main paper.

Ablation on DRN. We use Monodepth2 [2] to predict depth map in our framework, and we propose a Depth Refinement Network to refine the predicted depth map with uncertainty. Tab. 1 shows the ablation on our proposed DRN.

DRN	LM-O	YCB-V
×	57.2	58.3
✓	**58.7**	**60.6**

Table 1. Ablation on DRN. We report ADD(-S) on LM-O and YCB-V datasets here.

2. B. Running time

All our experiments are implemented using PyTorch [4]. We test our method on a PC with an Intel E5-2630 CPU and a GTX 3090 GPU. Given a 640×480 image, our approach takes $\approx$ 15 ms for correspondence extraction and $\approx$ 10 ms for 6D pose estimation.

3. C.More Results of DGECN.

In this section, we provide more detailed results on YCB-V dataset and qualitative results on YCB-V and LM-O datasets. We present detailed evaluation results on YCB-V [6] for our DGENC in Tab. 2 and we demonstrate additional qualitative results for LM-O [1] in Fig. 1. The evaluation protocol of BOP Challenge has recently become more popular. Therefore, as shown in Tab. 3, we also present the results of our DGECN on LM-O and YCB-V under the BOP setup.

*Chunxia Xiao and Fei Luo are co-corresponding authors

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014. 1

[2] Clement Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3827–3837, 2019. 1

[3] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3385–3394, 2019. 1

[4] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 1

[5] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on*

Method	PoseCNN	SegDriven	Single-Stage	GDR-Net	DGECN(**Ours**)
002 master chef can	3.6	33.0	-	41.5	**45.3**
003 cracker box	25.1	44.6	-	**83.2**	77.5
004 sugar box	40.3	75.6	-	91.5	**94.8**
005 tomato soup can	25.5	40.8	-	65.9	**71.2**
006 mustard bottle	61.9	70.6	-	**90.2**	89.9
007 tuna fish can	11.4	18.1	-	44.2	**54.3**
008 pudding box	14.5	12.2	-	2.8	**16.7**
009 gelatin box	12.1	59.4	-	61.7	**62.2**
010 potted meat can	18.9	33.3	-	64.9	**65.8**
011 banana	30.3	16.6	-	64.1	**78.9**
019 pitcher base	15.6	90.0	-	**99.0**	98.5
021 bleach cleanser	21.2	70.9	-	73.8	**82.1**
024 bowl^S	12.1	30.5	-	**37.7**	23.5
025 mug	5.2	40.7	-	61.5	**63.5**
035 power drill	29.9	63.5	-	**78.5**	77.2
036 wood block^S	10.7	27.7	-	59.5	**62.3**
037 scissors	2.2	17.1	-	3.9	**18.3**
040 large marker	3.4	4.8	-	7.4	**8.1**
051 large clamp^S	28.5	25.6	-	**69.8**	55.6
052 extra large clamp^R	19.6	8.8	-	90.0	**90.1**
061 foam brick^S	54.5	34.7	-	**71.9**	38.6
Average	21.3	39.0	53.9	60.1	**60.6**

Table 2. Detailed results on YCB-V w.r.t. ADD(-S). (S) denotes symmetric objects.

Method	Ref.	**LMO**			**YCB-V**			Mean AR
		AR_{VSD}	AR_{MSSD}	AR_{MSPD}	AR_{VSD}	AR_{MSSD}	AR_{MSPD}	
CosyPose	✓	**0.480**	0.606	0.812	**0.772**	**0.842**	**0.850**	**0.727**
EPOS		0.389	0.501	0.750	0.626	0.677	0.783	0.621
PVNet		0.428	0.543	0.754	-	-	-	-
CDPN		0.445	**0.612**	0.815	0.396	0.570	0.631	0.578
GDR-Net		-	-	-	0584	0.674	0.726	-
SO-Pose		0.442	0.581	**0.817**	0.652	0.731	0.763	0.664
Ours		0.458	0.593	0.816	0.663	0.726	0.775	0.672

Table 3. Comparison with state-of-the-art methods on LMO and YCB-V under BOP metrics. We provide results for AR_{VSD}, AR_{MSSD} and AR_{MSPD} on LMO and YCB-V. Mean AR represents the overall performance on these two datasets as the average over all AR scores. Overall best results are in bold.

Computer Vision and Pattern Recognition, pages 4561–4570, 2019. 1

[6] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1

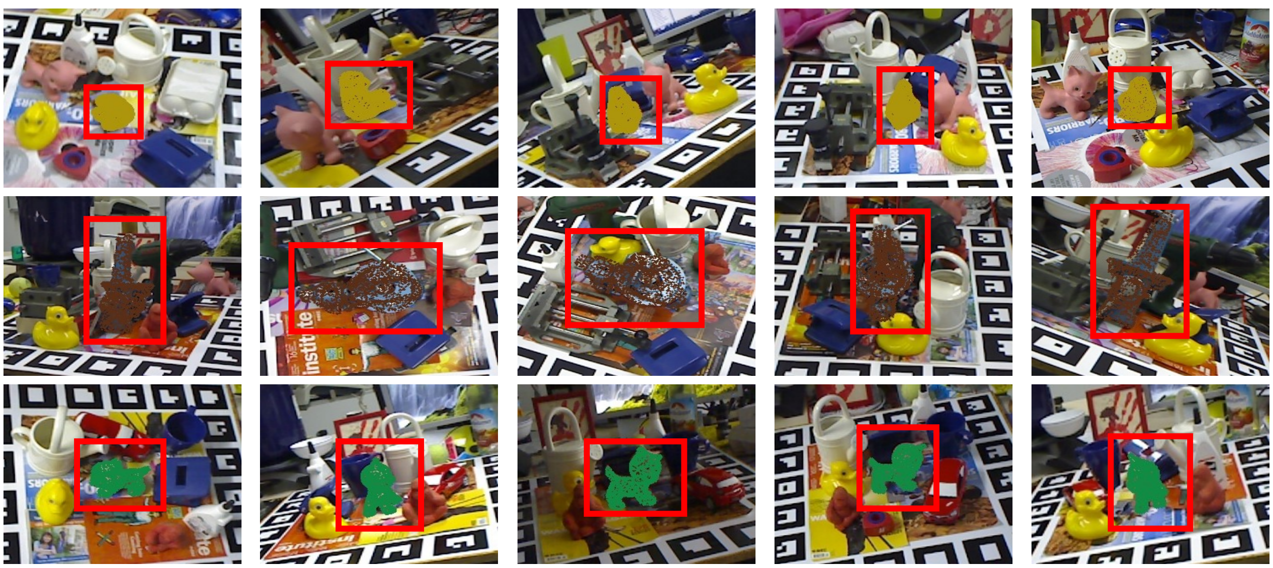

Figure 1. **Qualitative results on LM-O.** Here, the pose is visualized as the reprojection of the 3D mesh for each object.