Learning Better Keypoints for Multi-Object 6DoF Pose Estimation

Yangzheng Wu , Michael Greenspan
RCV Lab, Dept. of Electrical and Computer Engineering, Ingenuity Labs,
Queen's University, Kingston, Ontario, Canada
{y.wu, michael.greenspan}@queensu.ca

S.1. Overview

We document here the network structure, some additional results, and one more ablation study. The network diagram is shown in Figure. S.1. It is directly taken from the classification structure of edge-conv [2] with a few changes to the number of intermediate channels and the shape of the output vector. The per category KeyGNet results of LMO and YCB-V datasets evaluated by ADD(S) and ADD(S) AUC in both SISO and MIMO modes on all three methods tested are shown in Table. S.2, Table. S.3, and Table. S.4. KeyGNet keypoints improved the perfomance for all objects, in all datasets, among all three methods tested. The BOP AR (Average Recall) of Visible Surface Discrepancy (AR_{VSD}), Maximum Symmetry-Aware Surface Distance (AR_{MSSD}), Maximum Symmetry-Aware Projection Distance (AR_{MSPD}), and the overall average are reported in Table. S.5. All these metrics are improved in all six core datasets when the KeyGNet keypoints are used. Last but not least, the SISO-MIMO gap is reduced by using KeyGNet keypoints for all objects in YCB-V, as shown in Table. S.6.

S.2. Classical Distance Measure vs. KeyGNet

Instead of training a network, keypoints can be selected by measuring the Wasserstein distance directlyon a collection of sets of keypoints. We conduct a test by comparing the trained KeyGNet with a classical RANSAC [1] style algorithm. The collection of initial keypoint sets are selected either relatively randomly in a region centered at the bounding box's corners, or completely randomly in a sphere within the object reference frame of the CAD model. The Wasserstein distances and the dispersion scores are then calculated for each set of the keypoints. The algorithm repeats for N times and the keypoints with the minimum Wasserstein distances and the maximum dispersion scores are selected.

We test the keypoints using RCVPose on LMO and compare the ADD(S) metric with KeyGNet. The results are shown in Table. S.1. It can be seen that he keypoints selected with an initial location of bounding box corners are 3.2% on average worse than those selected with completely random initial locations. This is possibly due to the restrictions caused by the initial input locations of BBox corners. The learned KeyGNet keypoints have the best performance for all objects in LMO, boosting the ADD(S) by 13.6% and 10.4% compared to those randomly selected.

LMO Object	Random		KGN
	BBox	Sphere	
ape	53.7	55.2	**65**
can	80.8	83.2	**96.4**
cat	44.1	47.3	**58**
driller	70.6	73.4	**82.1**
duck	42.1	48.2	**65.6**
eggbox	70.1	74.3	**82.2**
glue	66.2	67.3	**75.1**
holepuncher	68.5	72.5	**81.2**
average	62	65.2	**75.6**

Table S.1. ADD(S) of RCVPose [3] on LMO using keypoints selected randomly (BBox, Sphere) vs. with KeyGNet (KGN). The randomly selected keypoints use RANSAC to minimize the Wasserstein distance measure.

References

[1] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 1

[2] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 1, 2

[3] Yangzheng Wu, Mohsen Zand, Ali Etemad, and Michael Greenspan. Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 335–352. Springer, 2022. 1

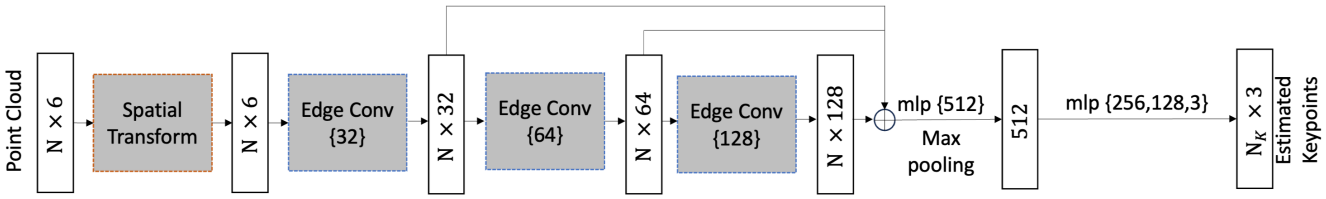

Figure S.1. KeyGNet Network Structure. The network is based on the classification structure of edge-conv [2]. The spatial transform block and edge-conv blocks are exactly the same as in the original setup. The output vector is reshaped to $N_K \times 3$ keypoints.

	SISO						MIMO					
LMO	PVNet		PVN3D		RCVPose		PVNet		PVN3D		RCVPose	
object	FPS	KGN	FPS	KGN	BBox	KGN	FPS	KGN	FPS	KGN	BBox	KGN
ape	15.8	21.2	33.9	40.2	61.3	**65.4**	8.2	20.9	25.7	39.9	57.1	65
can	63.3	74.2	88.6	93.7	93	**96.4**	51.1	74.2	77.4	93.7	90.1	**96.4**
cat	16.7	22.3	39.1	49.2	51.2	**58.2**	9.6	22.3	31.5	48.9	45.5	58
driller	65.7	76.6	78.4	88.3	78.8	81.7	57.5	76.6	68.1	88.3	77.5	**82.1**
duck	25.2	30.2	41.9	47.6	53.4	65.2	14.1	29.9	32.2	47.2	46.2	**65.6**
eggbox	50.2	57.8	80.9	85.2	82.3	82	38	57.8	70.7	85.2	80.2	**82.2**
glue	49.6	59.7	68.1	77.2	72.9	74.9	42.9	59.7	58.3	76.8	66.6	**75.1**
holepuncher	39.7	42.3	74.7	82.3	75.8	81.2	32.4	42.3	67.5	**82.3**	73.7	81.2
average	40.8	48 (+7.2)	63.2	70.5 (+7.3)	71.1	75.5 (+4.4)	31.7	47.9 (**+16.2**)	53.9	70.3	64.8	75.6 (+10.8)

Table S.2. LMO Results: The ADD(S) AUC comparison of three keypoint voting-based methods (PVNet, PVN3D, RCVPose) using initially defined keypoints (FPS, BBox) and optimized keypoints (KGN) generated by KeyGNet.

YCB	PVNet		PVN3D		RCVPose	
object	FPS	KGN	FPS	KGN	BBox	KGN
002_master_chef_can	54.6	69.5	75.3	84.8	92.1	**95.1**
003_cracker_box	66.2	78.8	87.0	**96.5**	94.3	96.4
004_sugar_box	66.3	76.1	90.9	96.5	94.2	**97.7**
005_tomato_soupcan	62.2	74.7	81.5	92.8	91.5	**96.7**
006_mustard_bottle	67.6	82.4	89.3	**97.7**	94.2	96.9
007Auna_fish_can	64.9	77.0	87.0	95.7	94.2	**96.4**
008_pudding_box	76.6	85.2	90.3	**97.5**	95.4	97.0
009_gelatin_box	71.4	88.8	90.6	**97.5**	92.6	97.4
010_potted_meat_can	69.5	84.1	82.5	93.7	88.1	**93.9**
011_banana	67.9	76.8	90.9	97.2	94.9	**97.5**
019_pitcher_base	67.8	76.9	88.1	97.5	93.1	**97.9**
021_bleach_cleanser	70.2	74.9	92.2	96.6	95.4	**98.4**
024_bowl*	66.9	78.3	87.3	95.6	91.0	**97.3**
025_mug	71.5	79.8	91.8	96.5	94.2	**96.9**
035_power_drill	67.6	81.7	89.2	96.9	93.7	**97.5**
036_wood_block*	57.4	85.2	82.9	92.8	89.7	**93.3**
037_scissors	64.2	80.4	83.2	91.5	92.3	**95.9**
040_large_marker	65.5	81.9	84.2	88.0	86.5	**95.6**
051_large_clamp*	55.7	66.5	84.4	88.5	90.5	**97.6**
052_extra_large_clamp*	52.6	61.4	77.8	94.1	92.5	**96.0**
061_foam_brick*	60.9	79.6	89.4	**97.8**	92.2	97.2
average	65.1	78.1	86.5	94.6	92.5	**96.6**

Table S.3. YCB-V Results. The ADD(S) AUC comparison of three keypoint voting-based methods (PVNet, PVN3D, RCVPose) in MIMO mode using initially defined keypoints (FPS, BBox) and optimized keypoints (KGN) generated by KeyGNet.

YCB	PVNet		PVN3D		RCVPose	
object	FPS	KGN	FPS	KGN	BBox	KGN
002_master_chef_can	60.2	70	79.3	85.2	94.7	**96.2**
003_cracker_box	70.7	79.4	91.5	96.7	96.4	**97.4**
004_sugar_box	73.2	76.6	96.9	97.3	97.6	**98.7**
005_tomato_soupcan	67.7	75.1	89.0	93.2	95.4	**97.6**
006_mustard_bottle	76.5	83	97.9	**98.2**	97.7	**98.2**
007Auna_fish_can	71.3	77.2	90.7	96.3	96.7	**97.4**
008_pudding_box	80.1	85.4	97.1	**98.1**	97.4	97.9
009_gelatin_box	81.2	89.1	**98.3**	**98.3**	97.9	**98.3**
010_potted_meat_can	76.9	84.6	87.9	94.2	92.6	**95.3**
011_banana	73.2	77.6	96.0	97.6	97.2	**98.4**
019_pitcher_base	74.3	77.4	96.9	98.0	96.7	**99.2**
021_bleach_cleanser	70.9	75.4	95.9	97.3	98.4	**99.3**
024_bowl*	69.7	79	92.8	96.4	95.3	**98.2**
025_mug	75.3	80.6	96.0	97.1	97.1	**98**
035_power_drill	74.3	82	95.7	97.2	96.9	**98.3**
036_wood_block*	70.2	85.8	91.1	93.2	90.7	**94.3**
037_scissors	66.4	81	87.2	92.1	94.9	**97.2**
040_large_marker	67.3	82.4	91.6	94.3	93.2	**96.3**
051_large_clamp*	66.2	72.2	95.6	96.2	96.2	**98.3**
052_extra_large_clamp*	63.4	66.9	90.5	94.7	95.1	**97.2**
061_loam_brick*	70.2	80.3	98.2	**98.4**	96.6	98.2
average	73.4	79.1	92.3	95.7	95.9	**97.6**

Table S.4. YCB-V Results. The ADD(S) AUC comparison of three keypoint voting-based methods (PVNet, PVN3D, RCVPose) in SISO mode using initially defined keypoints (FPS, BBox) and optimized keypoints (KGN) generated by KeyGNet.

Metric	Dataset	PVNet			PVN3D			RCVPose		
		FPS	BBox	KGN	FPS	BBox	KGN	FPS	BBox	KGN
AR_{VSD}	LMO	48.2	40.2	52.4	70.6	65.3	72.7	71.7	72.5	**76.9**
	YCB-V	78.2	70.4	82.7	76.9	75.4	83.6	83.7	84.4	**88.3**
	TLESS	65.7	58.7	67.2	68.3	66.7	72.2	70.4	70.8	**75.3**
	TUDL	90.5	85.5	92.5	87.3	86.2	91.9	97.2	98.0	**99.4**
	IC-BIN	70.6	64.3	72.6	67.2	63.4	73.9	73.7	74.1	**80.4**
	ITODO	42.4	33.5	44.7	43.2	41.7	47.7	48.2	50.7	**50.8**
	HB	77.1	74.3	78.7	78.4	75.6	84.4	81.6	82.5	**85.9**
AR_{MSSD}	LMO	66.4	60.2	70.2	62.5	61.3	69.3	72.7	73.4	**73.6**
	YCB-V	77.3	68.7	79.8	79.9	77.4	82.9	83.8	86.3	**89.6**
	TLESS	70.2	64.7	72.0	64.3	62.3	66.5	70.8	72.3	**73.3**
	TUDL	90.6	84.7	91.8	91.9	88.9	93.1	96.7	97.5	**98.5**
	IC-BIN	69.0	60.2	71.5	72.1	70.9	**78.4**	73.2	73.9	73.8
	ITODO	51.2	43.3	51.3	52.6	51.7	58.3	56.2	57.2	**64.2**
	HB	85.5	82.3	90.2	85.3	84.2	88.2	88.6	89.0	**90.4**
AR_{MSPD}	LMO	69.2	62.3	71.7	66.0	64.3	67.7	73.7	74.9	**79.7**
	YCB-V	76.7	72.1	77.0	76.5	75.4	81.2	83.7	84.9	**88.1**
	TLESS	67.3	60.2	71.1	69.3	67.8	72.1	70.2	71.5	**77.7**
	TUDL	93.7	90.2	94.9	91.2	90.2	95.1	96.7	97.8	**98.6**
	IC-BIN	72.3	64.3	75.4	71.8	70.2	**77.6**	72.7	73.9	75.6
	ITODO	50.3	42.7	52.3	52.7	51.3	55.6	55.6	56.2	**59.4**
	HB	84.8	81.2	86.0	84.7	83.4	90.2	89.2	90.5	**93.0**
$AR_{average}$	LMO	61.3	54.2	64.8	66.4	63.6	69.9	72.7	73.6	**76.7**
	YCB-V	77.4	70.4	79.8	77.8	76.1	82.6	83.7	85.2	**88.7**
	TLESS	67.7	61.2	70.1	67.3	65.6	70.3	70.5	71.5	**75.4**
	TUDL	91.6	86.8	93.1	90.1	88.4	93.4	96.9	97.8	**98.8**
	IC-BIN	70.6	62.9	73.2	68.2	70.4	**76.6**	73.2	74.0	**76.6**
	ITODO	48.0	39.8	49.4	48.2	49.5	53.9	53.3	54.7	**58.1**
	HB	82.5	79.3	85.0	82.8	81.1	87.6	86.5	87.3	**89.7**
Overall Average		71.3	64.9	73.6	72.0	70.7	76.3	76.7	77	**80.6**

Table S.5. BOP Core Dataset Results. The Average Recall (AR) of Visible Surface Discrepancy (AR_{VSD}), Maximum Symmetry-Aware Surface Distance (AR_{MSSD}), Maximum Symmetry-Aware Projection Distance (AR_{MSPD}), and the overall average for all six BOP core datasets are reported for three methods by using keypoints selected by the original method (FPS/BBox) and KeyGNet (KGN).

YCB-V	PVNet		PVN3D		RCVPose	
object	FPS	KGN	FPS	KGN	BBox	KGN
002_master_chef_can	-5.6	-0.5	-4.0	**-0.4**	-2.6	-1.1
003_cracker_box	-4.5	-0.6	-4.5	**-0.2**	-2.1	-1.0
004_sugar_box	-6.9	**-0.5**	-6.0	-0.8	-3.4	-1.0
005_tomato_soupcan	-5.5	**-0.4**	-7.5	**-0.4**	-3.9	-0.9
006_mustard_bottle	-8.9	-0.6	-8.6	**-0.5**	-3.5	-1.3
007Auna_fish_can	-6.4	**-0.2**	-3.7	-0.6	-2.5	-1.0
008_pudding_box	-3.5	**-0.2**	-6.8	-0.6	-2.0	-0.9
009_gelatin_box	-9.8	**-0.3**	-7.7	-0.8	-5.3	-0.9
010_potted_meat_can	-7.4	**-0.5**	-5.4	**-0.5**	-4.5	-1.4
011_banana	-5.3	-0.8	-5.1	**-0.4**	-2.3	-0.9
019_pitcher_base	-6.5	**-0.5**	-8.8	**-0.5**	-3.6	-1.3
021_bleach_cleanser	-0.7	**-0.5**	-3.7	-0.7	-3.0	-0.9
024_bowl*	-2.8	**-0.7**	-5.5	-0.8	-4.3	-0.9
025_mug	-3.8	-0.8	-4.2	**-0.6**	-2.9	-1.1
035_power_drill	-6.7	**-0.3**	-6.5	**-0.3**	-3.2	-0.8
036_wood_block*	-12.8	-0.6	-8.2	**-0.4**	-1.0	-1.0
037_scissors	-2.2	**-0.6**	-4.0	**-0.6**	-2.6	-1.3
040_large_marker	-1.8	**-0.5**	-7.4	-6.3	-6.7	-0.7
051_large_clamp*	-10.5	-5.7	-11.2	-7.7	-5.7	**-0.7**
052_extra_large_clamp*	-10.8	-5.5	-12.7	**-0.6**	-2.6	-1.2
061_loam_brick*	-9.3	-0.7	-8.8	**-0.6**	-4.4	-1.0
average	-6.3	**-1.0**	-6.7	-1.2	-3.4	**-1.0**

Table S.6. SISO-MIMO performance gap on YCB-V. The change in ADD(S) when converting from SISO to MIMO, for keypoints sampled heuristically (FPS or BBox) and KeyGNet (KGN). There is a relatively small change in ADD(S) AUC when the PE network is trained simultaneously on multiple objects using KGN keypoints.