Supplementary Material - RDPN6D: Residual-based Dense Point-wise Network for 6Dof Object Pose Estimation Based on RGB-D Images

Zong-Wei Hong Yen-Yang Hung Chu-Song Chen
National Taiwan University
{r10922190, r11922a18}@ntu.edu.tw, chusong@csie.ntu.edu.tw

A. More Implementation Details

More implementation details are provided in this section.

A.1. Network Architecture

The detailed architecture of the proposed RDPN is shown in Fig. 1. In this figure, *conv*$(n * n, c)$ denotes a 2D convolution with kernel size n and output channel c. bn denotes batch normalization, $relu$ denotes ReLU activation, *Upsample*(s) denotes 2D upsampling with scale factor s. and *maxpool*(k, s, p) denotes 2D max pooling with kernel size k, stride s, and padding p, respectively. The output of *adaptive avgpool*(h, w) or *adaptive maxpool*(h, w) is of size $h * w$ for any input size. *convTranspose*$(n * n, c)$ denotes a 2D transposed convolution with kernel size n and output channel c. gn denotes group normalization [21], $Leakyrelu$ denotes LeakyReLU activation, and *Linear*(c) denotes a fully connected layer with output channel c.

To represent rotations, we adopt the solution proposed in [28] to address the issue of rotation discontinuity, which results in a 6-dimensional output.

A.2. Training Parameters

For RDPN, all networks were trained using the Ranger optimizer [13,26] with a batch size of 24 and an initial learning rate of 1e-4. This learning rate was gradually reduced using a cosine schedule [14] at 72% of the training process.

A.3. Training Enhancements

We employ two strategies to enhance the model's ability to handle objects of varying sizes. First, we dynamically adjust the receptive field of the $\mathcal{F}_{residual}$ based on the size of the corresponding tight 3D bounding box of the CAD model. This allows the model to focus more effectively on objects of different scales.

Second, we adopt the Dynamic Zoom-In technique proposed in [10, 20] to alleviate the impact of varying object sizes further. During training, we randomly shift the center and scale of the ground-truth bounding boxes by a ratio of 25%. Subsequently, we zoom in the input Regions of Interest (RoIs) with a ratio of r = 1.5 while maintaining their original aspect ratio. This ensures that the area containing the object occupies approximately half of the RoIs. This dynamic zooming approach effectively normalizes the object size distribution and improves the model's generalization ability across different object sizes.

	DenseFusion [19]	FFB6D [4]	ES6D [15]	RDPN (Ours)
ADD-S	93.2	95.0	93.6	**95.4**
ADD(-S)	86.1	91.3	89.0	**91.5**

Table 1. **The YCB-V results with PoseCNN input.**

B. More Results

This section presents detailed evaluations of RDPN on the MP6D, YCB-Video datasets, and the BOP challenge [6].

B.1. Quantitative Results under the same detections on the YCB-V Dataset

To comprehensively assess the effectiveness of RDPN, we compare it with several baseline methods while ensuring a fair comparison. However, it is essential to note that while other methods utilize segmentation masks or built-in detection techniques, RDPN incorporates detection preprocessing specifically designed for RGBD images. Therefore, we adopt PoseCNN's [24] RoI results for RDPN and segmentations for other methods to maintain consistency and impartiality. Despite this disparity in detection pipelines, RDPN exhibits robust accuracy, as evidenced in Tab. 1. This finding underscores its efficacy even when operating under different detection paradigms.

B.2. Quantitative Results on the BOP challenge

Tab. 2 presents the average recall for the BOP challenge, a comprehensive benchmark for rigid body pose estimation encompassing seven diverse datasets. This benchmark has

Method	Refinement	LM-O	T-LESS	TUD-L	YCB-V	ITODD	HB	IC-BIN	Avg(7)
RCVPose 3D_SingleModel_VIVO_PBR [22] (3DV' 22)	ICP	0.729	0.708	**0.966**	0.843	0.536	0.863	**0.733**	0.768
SurfEmb-PBR-RGBD [3] (CVPR' 22)	custom	0.760	0.828	0.854	0.799	0.538	0.866	0.659	0.758
ZebraPoseSAT-EffnetB4_refined [17] (CVPR' 22)	CIR [12] (CVPR' 22)	0.780	**0.862**	0.956	**0.899**	**0.618**	**0.921**	0.654	**0.813**
RADet+PFA-MixPBR-RGBD [2] (CVPR' 23)	PFA [8] (ECCV' 22)	**0.797**	0.850	0.960	0.888	0.469	0.869	0.676	0.787
RDPN (Ours)	CIR [12] (CVPR' 22)	0.776	0.768	0.957	0.883	0.575	0.907	0.720	0.798

Table 2. **Average Recall on the BOP Core datasets.**

Method	Hodan [7]	PointFusion [25]	DCF [11]	DF (per-pixel) [19]	MaskedFusion [16]	G2L-Net [1]	PVN3D [5]	FFB6D [4]	DFTr [27]	RDPN (Ours)
Obj_01	83.42	84.33	86.06	89.35	88.95	89.51	90.28	93.28	95.44	**99.58**
Obj_02	80.23	81.01	85.36	87.78	89.19	89.03	91.88	92.83	96.51	**99.19**
Obj_03	65.78	64.74	65.33	72.45	70.03	74.93	76.67	79.51	84.93	**93.87**
Obj_04	70.56	72.50	73.95	77.98	74.68	85.39	88.13	84.98	92.02	**96.36**
Obj_05	69.78	68.96	67.19	71.23	75.69	72.13	73.46	76.33	86.24	**95.30**
Obj_06	72.36	70.66	71.65	75.34	78.31	85.08	87.16	83.98	96.10	**96.30**
Obj_07	80.79	81.12	82.07	88.63	85.25	89.09	94.81	94.94	97.51	**99.33**
Obj_08	80.71	81.37	82.39	84.78	85.38	90.10	93.76	89.76	96.75	**99.23**
Obj_09	69.80	65.98	68.27	73.67	75.46	79.91	82.71	81.25	91.23	**95.00**
Obj_10	75.32	77.19	79.10	80.54	77.62	86.03	86.16	88.92	94.98	**98.34**
Obj_11	72.56	71.98	70.96	79.65	75.91	82.01	81.21	84.87	92.36	**92.55**
Obj_12	74.13	76.32	77.03	78.88	76.98	77.93	79.00	84.82	89.99	**95.89**
Obj_13	78.63	77.05	75.15	80.12	80.58	85.38	86.69	85.42	95.04	**95.80**
Obj_14	76.89	77.90	76.98	80.89	81.15	84.54	87.06	87.99	94.13	**94.87**
Obj_15	64.53	67.36	66.23	68.45	66.30	72.92	74.17	75.01	86.97	**88.90**
Obj_16	69.88	72.28	73.08	75.81	73.86	79.38	81.35	83.95	92.14	**94.93**
Obj_17	77.42	85.93	84.68	89.16	88.11	92.08	93.47	93.19	94.25	**98.83**
Obj_18	75.63	81.46	80.91	83.23	85.94	88.13	87.57	91.73	94.69	**98.13**
Obj_19	72.89	76.82	78.07	81.98	79.37	85.31	88.82	87.28	95.03	**96.13**
Obj_20	72.65	75.91	74.20	76.59	78.93	81.41	88.10	85.75	**93.92**	89.56
Avg (20)	74.20	75.54	75.93	79.84	79.38	83.51	85.42	86.29	93.01	**95.90**

Table 3. **Quantitative evaluation of 6D Pose ADD-S AUC on the MP6D Dataset for each object.** Note that all objects are symmetric.

Method	Pre-process	Network	Post-process	Network + Post-process
DenseFusion [19]	*IS*	50	11	61
FFB6D [4]	-	42	65	107
ES6D [15]	*IS*	6	-	6
Uni6D* [9]	-	39	-	39
Uni6Dv2* [18]	-	47	-	47
RCVPose [23]	-	50	-	50
RDPN (Ours)	*OD*	20	-	20

Table 4. **Time Costs (in milliseconds per frame) on the YCB-Video Dataset.** *IS* represents Instance Segmentation, and *OD* represents Object Detection. (*) stands for methods whose source codes have not been released, and we report their speeds directly from their respective papers.

yet to reach saturation, indicating its suitability for evaluating the generalizability of pose estimation models. We evaluate RDPN on this challenge and compare its performance with published works.

B.3. Quantitative Results on the MP6D Dataset

The results of ADD-S AUC of each object on the MP6D dataset are shown in Tab. 3.

B.4. Time Costs Comparison on YCB-Video Dataset

The time costs comparison on YCB-Video dataset are shown in Tab. 4.

B.5. Visualization on Predicted Pose on the YCB-Video and MP6D Datasets

We provide several qualitative comparison results between our method and the previous state-of-the-art method FFB6D [4] in Fig. 2 for the YCB-Video dataset. Additionally, we provide several qualitative results on the MP6D dataset in Fig. 3.

The results demonstrate the effectiveness of our method on both datasets, including *texture-less* and *high-reflectivity* objects.

References

[1] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4233–4242, 2020. 2

[2] Yang Hai, Rui Song, Jiaojiao Li, Mathieu Salzmann, and Yinlin Hu. Rigidity-aware detection for 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8927–8936, 2023. 2

[3] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6749–6758, 2022. 2

[4] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3003–3013, 2021. 1, 2, 6

[5] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020. 2

[6] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020. 1

[7] Tomáš Hodaň, Xenophon Zabulis, Manolis Lourakis, Štěpán Obdržálek, and Jiří Matas. Detection and fine 3d pose estimation of texture-less objects in rgb-d images. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4421–4428. IEEE, 2015. 2

[8] Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Perspective flow aggregation for data-limited 6d object pose estimation. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[9] Xiaoke Jiang, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, and Liwei Wu. Uni6d: A unified cnn framework without projection breakdown for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11184, 2022. 2

[10] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019. 1

[11] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 641–656, 2018. 2

[12] Lahav Lipson, Zachary Teed, Ankit Goyal, and Jia Deng. Coupled iterative refinement for 6d multi-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6728–6737, 2022. 2

[13] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 1

[14] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[15] Ningkai Mo, Wanshui Gan, Naoto Yokoya, and Shifeng Chen. Es6d: A computation efficient and symmetry-aware 6d pose regression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6718–6727, 2022. 1, 2

[16] Nuno Pereira and Luís A Alexandre. Maskedfusion: Mask-based 6d object pose estimation. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 71–78. IEEE, 2020. 2

[17] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 2

[18] Mingshan Sun, Ye Zheng, Tianpeng Bao, Jianqiu Chen, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xiaoke Jiang. Uni6dv2: Noise elimination for 6d pose estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 1832–1844. PMLR, 2023. 2

[19] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 1, 2

[20] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 1

[21] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 1

[22] Yangzheng Wu, Alireza Javaheri, Mohsen Zand, and Michael Greenspan. Keypoint cascade voting for point cloud based 6dof pose estimation. In *2022 International Conference on 3D Vision (3DV)*, pages 176–186. IEEE, 2022. 2

[23] 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 *European Conference on Computer Vision*, pages 335–352. Springer, 2022. 2

[24] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018. 1

[25] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In

Proceedings of the IEEE conference on computer vision and pattern recognition, pages 244–253, 2018. 2

[26] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019. 1

[27] Jun Zhou, Kai Chen, Linlin Xu, Qi Dou, and Jing Qin. Deep fusion transformer network with weighted vector-wise keypoints voting for robust 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13967–13977, October 2023. 2

[28] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 1

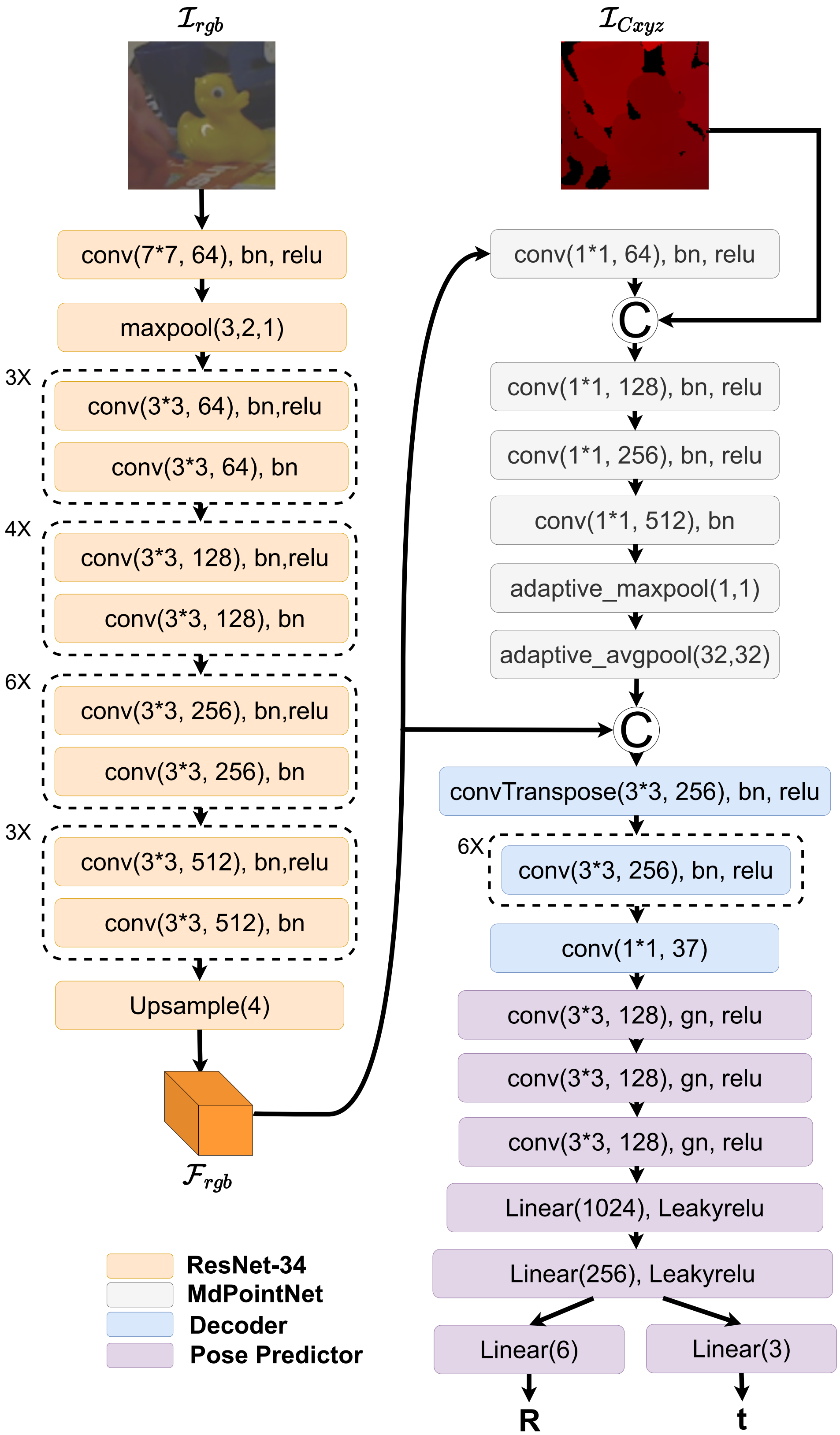

Figure 1. **The detailed architecture of our proposed RDPN framework.**

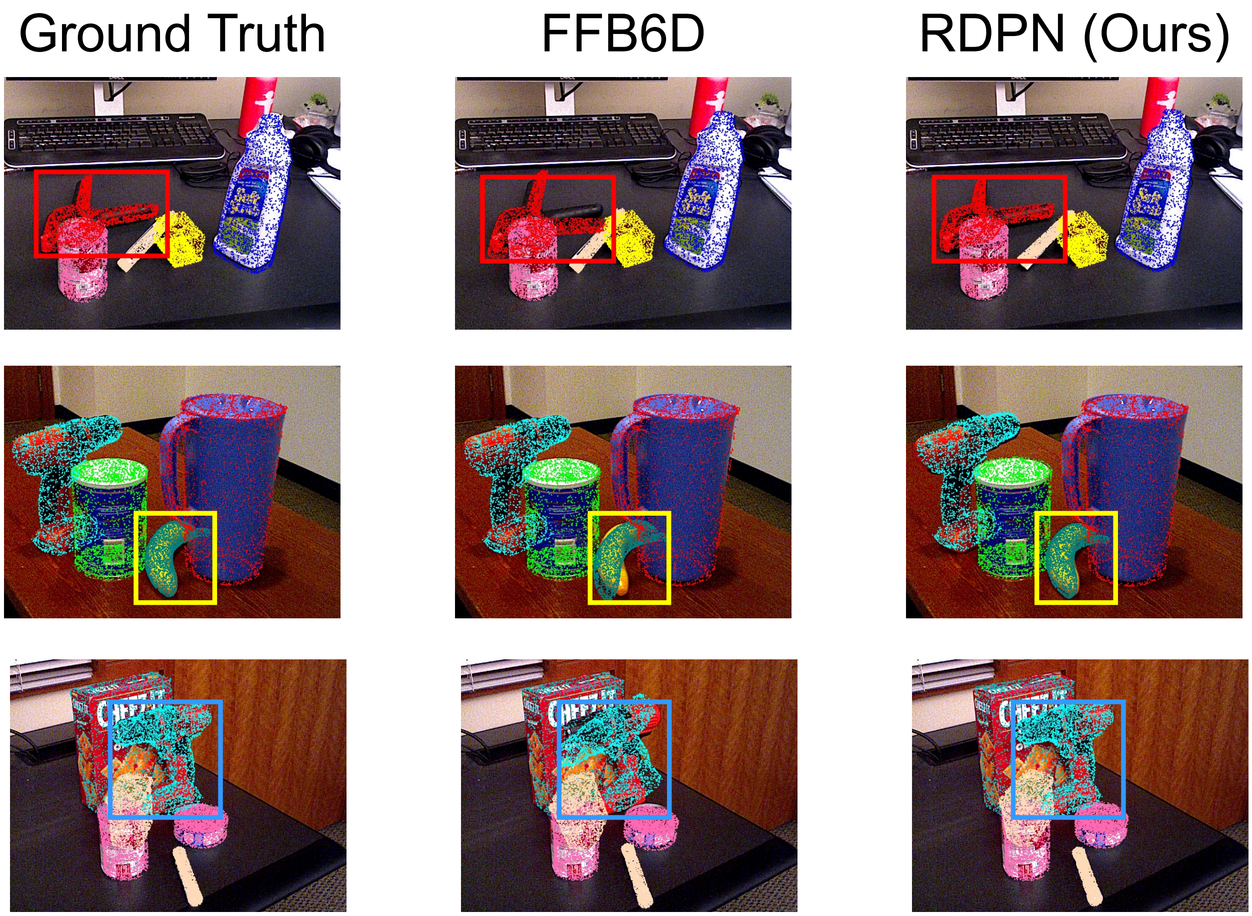

Figure 2. **Qualitative results on YCB-Video dataset.** The first column shows the ground truth pose. The second column shows the pose estimated using the keypoint-based method FFB6D [4]. The third column shows the pose estimated using our RDPN approach. Inside the bounding box, we see that our dense correspondence method outperforms the keypoint-based method FFB6D [4] in handling pose estimation under occlusion conditions.

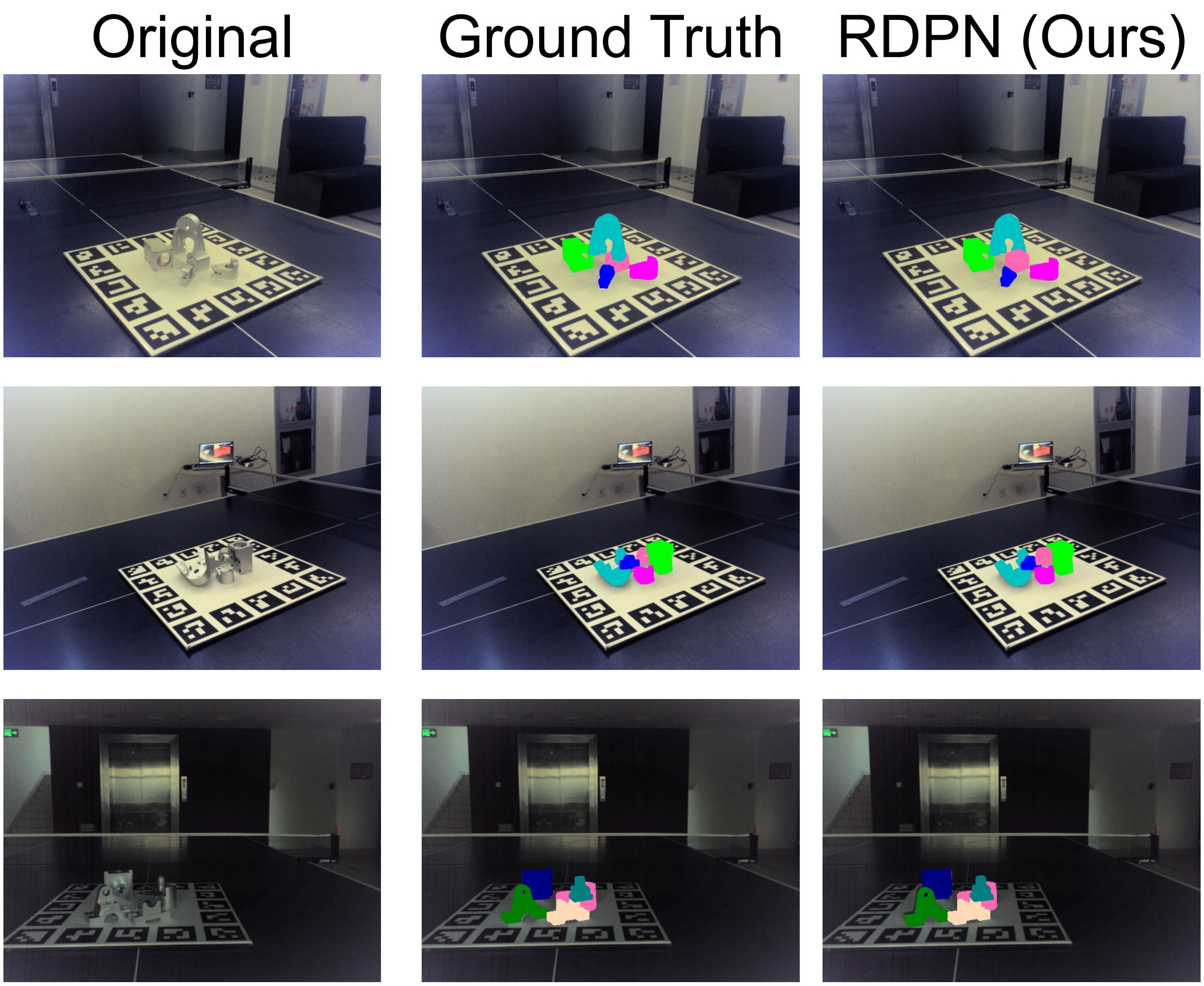

Figure 3. **The qualitative results on MP6D dataset.** All images are rendered by projecting the 3D object model onto the image plane using the estimated pose. The results demonstrate the effectiveness of our method on texture-less and high-reflectivity objects under various lighting conditions.