Supplementary Material for "DCL-Net: Deep Correspondence Learning Network for 6D Pose Estimation"

Hongyang Li, Jiehong Lin, and Kui Jia

A More Implementation Details of Point-wise Feature Extraction.

For point-wise feature extraction (cf. Sec. 3.1), we employ two backbones with the same architectures to capture point-wise feature maps $\boldsymbol{F}^{\mathcal{X}_c}$ and $\boldsymbol{F}^{\mathcal{Y}_o}$ from the object observation and its CAD model, respectively.

For each branch, we firstly quantify the point set of the input object, attached with RGB values, into $64 \times 64 \times 64$ voxels; point coordinates and RGB values of points within a same voxel are averaged, resulting in a 6-dimensional vector. The volumetric input with a size of $64 \times 64 \times 64 \times 6$ is then fed into the backbone, which is constructed based on 3D Sparse Convolutions [2]. Fig. 1 illustrates the detailed architecture of the backbone, where network specifics are also given. As shown in the figure, the backbone stacks 8 convolutional layers and 4 pooling layers, point-wise features are interpolated from the convolutional feature map via a Tensor-to-Point module [3]. To enrich the features, we aggregate multi-scale point-wise features from 4 intermediate feature maps as the outputs of the backbone.

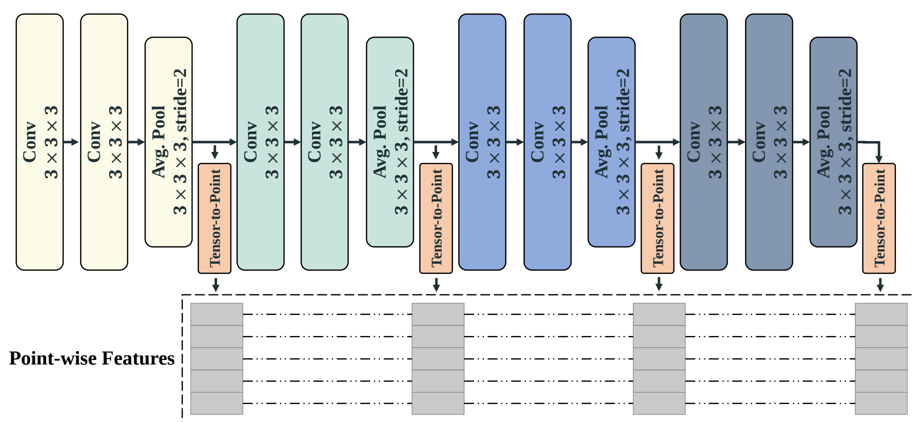

Fig. 1. An illustration of the architecture of backbone.

B More comparisons with other methods.

We report results on the metrics of both ADD-S AUC and ADD-S $<$ 2cm for YCB-Video dataset [1] to compare with the prior works [4–7]; however, those metrics *w.r.t* ADD-S are too relaxed to reflect the actual errors of poses, as verified in Fig. 2, where some predictions with small values of ADD-S/ADD(S), *e.g.*, ADD-S $<$ 2 cm, yet impose large pose errors to the ground truths. We thus include the results on the metric of $n°m$ cm, which denotes mean Average Precise (mAP) of objects with rotation error less than $n°$ and translation error less than m cm, in Table 1, and visualize the curves of Average Precision (AP) versus different thresholds of rotation and translation errors, respectively, both of which indicate that our DCL-Net outperforms the existing methods by a larger margin in the regime of high precision, especially the rotation estimation.

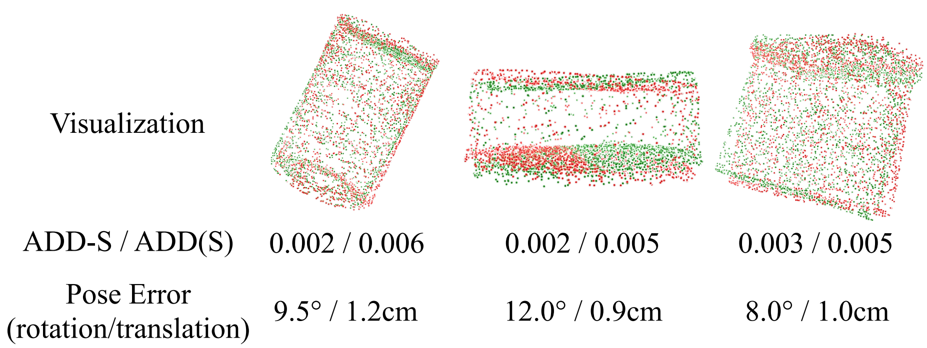

Fig. 2. Visualization of examples with small ADD-S / ADD(S) and large pose errors on YCB-Video dataset [1]. Point sets (green) denote object CAD models transformed by ground truth poses, while point sets (red) denote those transformed by the predicted ones.

Table 1. Quantitative comparisons on different evaluation metrics for YCB-Video dataset [1].

	DenseFusion [6]	PVN3D [5]	FFB6D [4]	DCL-Net
ADD-S AUC	93.1	95.5	**96.6**	**96.6**
ADD-S $<$ 2 cm	96.8	97.6	**99.2**	99.0
$2°2$ cm	19.4	14.6	22.8	**38.9**
$5°5$ cm	49.1	55.0	64.2	**65.2**

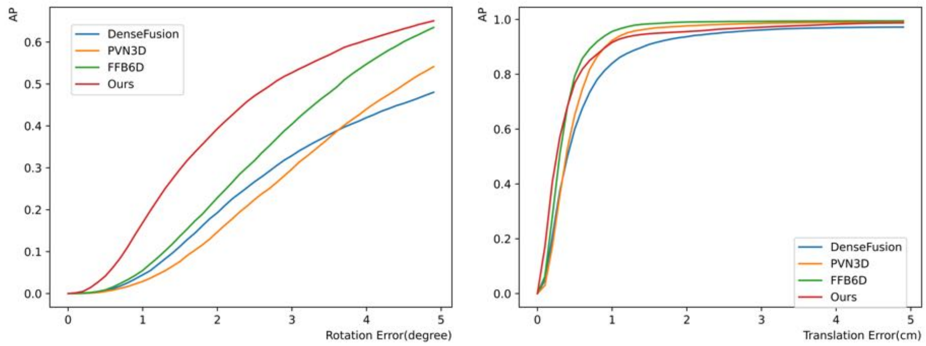

Fig. 3. Curves of average precision (AP) versus different thresholds of rotation and translation errors, respectively, on YCB-Video dataset [1].

References

1. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: 2015 international conference on advanced robotics (ICAR). pp. 510–517. IEEE (2015) 2, 3
2. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9224–9232 (2018) 1
3. He, C., Zeng, H., Huang, J., Hua, X.S., Zhang, L.: Structure aware single-stage 3d object detection from point cloud. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11873–11882 (2020) 1
4. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3003–3013 (2021) 2
5. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: 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. pp. 11632–11641 (2020) 2
6. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3343–3352 (2019) 2
7. Zhou, G., Wang, H., Chen, J., Huang, D.: Pr-gcn: A deep graph convolutional network with point refinement for 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2793–2802 (2021) 2