Prior-free Category-level Pose Estimation with Implicit Space Transformation – Supplementary

Outline

This is the supplementary material, which is divided into the following sections.

1. More implementation details are shown in Sec. 1
2. More experimental results are shown in Sec. 2
3. More visual comparisons are shown in Sec. 3
4. Limitation analysis and future work are in Sec. 4

1. More Implementation Details

1.1. Wild6D Dataset

Wild6D [3] contains 5,166 videos with 1722 object instances and 5 categories (bottle, bowl, camera, laptop, and mug). Among this data, 486 videos of 162 instances are split into the test set for model evaluation.

1.2. Training and Inference Details.

We train our IST-Net from scratch in an end-to-end manner for 30 epochs with a batch size of 24. We further employ the Adam optimizer with a base learning rate of 0.01. We adopt the StepLR scheduler with step size 1 and gamma as 5. Our experiments are conducted on two RTX3090Ti GPUs.

1.3. Network Configurations

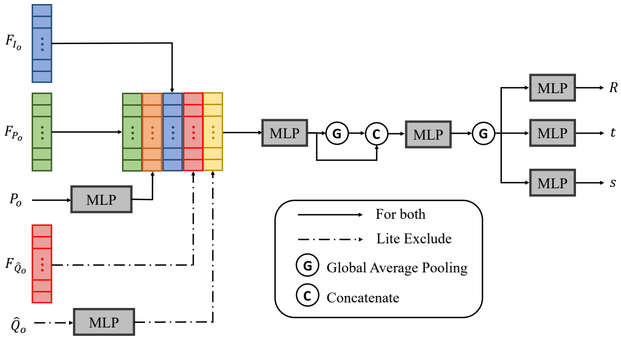

Figure 1. Architecture of pose estimators. The solid lines represent the same parts of all estimators, and the dashed line represents the part that is not adopted by the lite pose estimator.

As mentioned in the main paper, we provide the detailed architecture of the pose estimators, as shown in Fig. 1. IST-Net contains three pose estimators in camera-space enhancer, world-space enhancer, and final pose regression which follow similar architectures. The pose estimators in world-space enhancer and final pose regression share the same architecture and adopt a standard design, namely standard pose estimator. While the pose estimator in camera-space enhancer adopts a lightweight design, namely lite pose estimator. Specifically, in Fig. 1, the lite pose estimator only takes camera space information as input, including semantic features F_{P_o}, geometrical features F_{I_o} and position encoding term which is generated by MLP upon P_o. For the standard pose estimator, its inputs contain extra information from world-space, including world-space geometrical features $F_{\hat{Q}_o}$ and world-space position encoding term. Then the inputs are concatenated together and sent into an MLP to yield the fused features followed by a global average pooling layer. We further concatenate the global and local features and use a combination of MLP and a pooling layer to acquire the compressed features. Finally, three independent MLPs are used to predict R, t, and s respectively.

2. More Experimental Results

2.1. Results on CAMERA25 Dataset

We further report the results of our method on the CAMERA25 dataset, as shown in Tab. 1. Our method is competitive with other methods, specifically, on metric $3D_{75}$, IST-Net outperforms the previous state-of-the-art method by 2%. This indicates that our method has a strong ability to comprehensively estimate rotation, translation, and size.

2.2. Ablate on Shape Priors with Different Methods

In this part, we provide more experimental results to support the assumption “shape priors are not necessary” which is detailed in the main paper. We choose two competitive candidates from matching-based and regression-based methods, DPDN [4] and SGPA [1], using prior deformation. We list the experimental results in Tab. 2. We can find

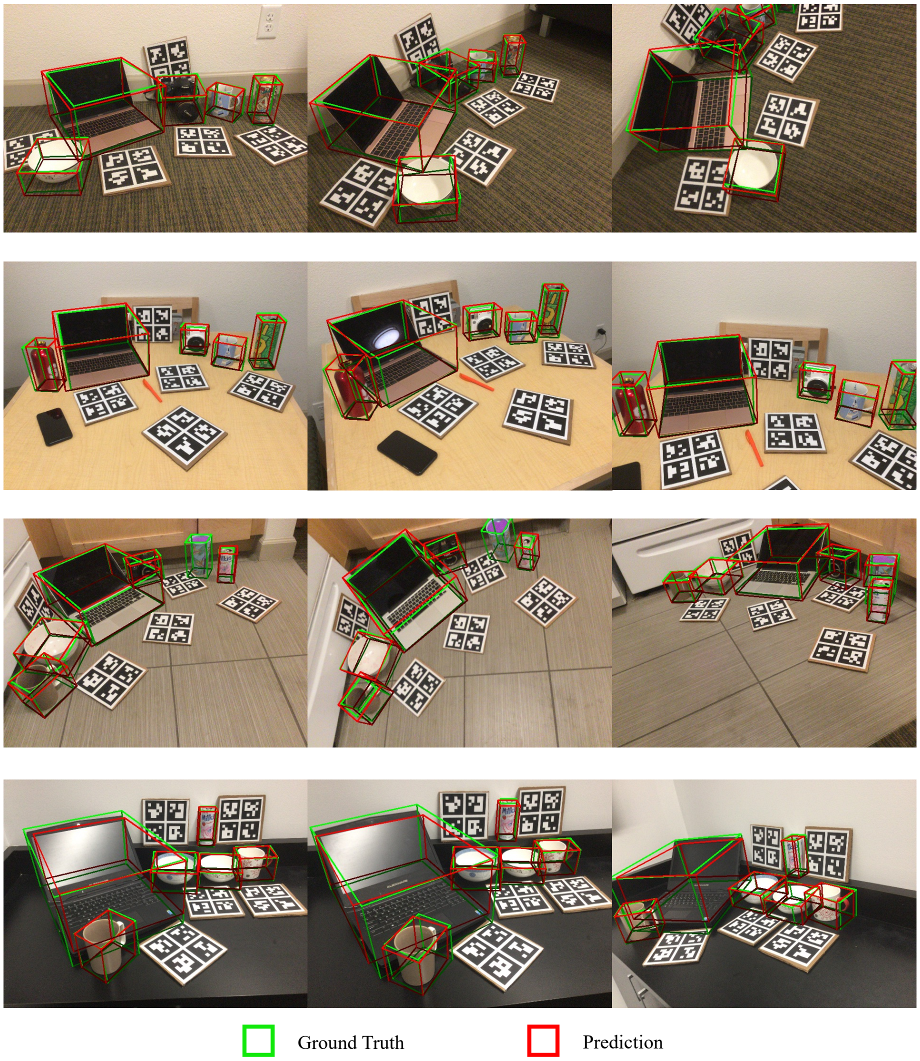

Figure 2. More visualization on REAL275 dataset.

Method	Prior	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$
NOCS [7]	✗	83.9	69.5	32.3	40.9	48.2	64.6
DualPoseNet [5]	✗	92.4	86.4	64.7	70.7	77.2	84.7
GPV-Pose [2]	✗	93.4	88.3	72.1	79.1	-	89.0
SPD [6]	✓	93.2	83.1	54.3	59.0	73.3	81.5
CR-Net [8]	✓	**93.8**	88.0	72.0	76.4	81.0	87.7
SGPA [1]	✓	93.2	88.1	70.7	74.5	**82.7**	88.4
RBP-Pose [9]	✓	93.1	89.0	**73.5**	79.6	82.1	89.5
IST-Net (Ours)	✗	93.7	**90.8**	71.3	**79.9**	79.4	**89.9**

Table 1. Comparison with state-of-art methods on CAMERA25 dataset. We summarize the pose estimation results reported in the original papers. **Prior** refers to whether the method builds upon shape priors. '-' denotes no results reported under this metric.

that regardless of whether the approach is a matching-based or a direct regression-based method when we use category-independent prior and noise to replace the default shape prior, the final performance does not have a significant difference. This phenomenon further reflects that shape prior is redundant for the prior deformation process, supporting our major claims in the main paper.

3. More Visualization

As shown in Fig. 2, we show more visualization of IST-Net on the REAL275 test split. As highlighted with the red box, ours can accurately predict the object pose, which visually demonstrates the superiority of our method.

4. Limitation Analysis and Future Work

Our method yields strong performance in NOCS and Wild6D datasets, but it might be sufficient for in-the-wild open-world evaluation, because, existing datasets contain limited object categories and the object structure is relatively simple.

Method	Prior	$3D_{25}$	$3D_{50}$	$3D_{75}$	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$10°10cm$
SGPA [1]	default	-	80.1	61.9	35.9	39.6	61.3	70.7	-
	bottle	83.9	81.0	65.5	37.0	42.1	58.6	69.9	-
	bowl	84.0	81.2	64.3	36.2	40.7	60.5	70.9	-
	camera	83.8	79.9	62.6	35.4	39.7	59.5	69.9	-
	can	84.1	80.8	65.1	36.5	41.5	59.3	70.4	-
	laptop	83.7	79.2	63.5	38.7	42.7	61.0	71.6	-
	mug	83.8	80.1	64.1	35.0	40.1	59.7	68.2	-
	noise	83.8	79.9	60.3	35.2	39.6	59.5	69.7	-
DPDNs [4]	default	84.2	83.4	76.0	46.0	50.7	70.4	78.4	80.4
	bottle	84.0	83.3	74.6	46.2	50.4	67.5	77.2	79.2
	bowl	83.8	83.2	75.9	46.1	51.3	68.0	78.1	80.1
	camera	84.0	82.3	73.5	45.5	53.1	66.9	77.9	80.1
	can	84.2	83.9	76.3	44.6	50.7	68.2	77.0	79.3
	laptop	83.4	81.4	73.2	44.2	49.2	67.9	77.2	79.9
	mug	84.1	84.0	76.6	45.9	50.3	68.9	77.4	79.7
	noise	84.2	83.8	76.1	45.7	51.0	69.5	77.7	79.8

Table 2. Ablate on shape priors with different Methods. ”default” represents the standard result obtained from the original paper. ‘-’ denotes no results are reported in the original literature.

We will work on building a category-level dataset with deiverse object types and shapes to further push forward the area. We hope our current investigation can shed light on more new insights in pose estimation.

References

[1] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021. 1, 2

[2] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6771–6781. IEEE, 2022. 2

[3] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *arXiv preprint arXiv:2206.15436*, 2022. 1

[4] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 19–34. Springer, 2022. 1, 2

[5] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. *CoRR*, abs/2103.06526, 2021. 2

[6] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 2

[7] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. 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*, pages 2642–2651, 2019. 2

[8] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. *CoRR*, abs/2108.08755, 2021. 2

[9] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 655–672. Springer, 2022. 2