Supplementary Material: Category-Level 6D Object Pose Estimation in the Wild: A Semi-Supervised Learning Approach and A New Dataset

Anonymous Author(s)
Affiliation
Address
`email`

A Discussion on Wild6D.

Statistics Analysis. We create a new large-scale benchmark for category-level object pose estimation called *Wild6D*, which consists of 5,166 videos (>1.1 million images) over 1722 different object instances and 5 categories, *i.e.*, *bottle*, *bowl*, *camera*, *laptop*, and *mug*. The distribution of the over objects per category in *Wild6D* is illustrated in Table 1.

Mask Segmentation Quality. Although the mask segmentation quality may influence the training process of RePoNet, *i.e.*, silhouette matching loss, we found in most cases the mask segmentations are satisfactory and good enough to conduct the silhouette matching loss. The mask segmentation results of several samples are shown in Fig.1. We provide the mask segmentation along with the raw RGB image and depth image for each video frame.

Baselines on Wild6D. In original paper, we evaluate several existing work on Wild6D testing set. More specifically, we utilize the official released model trained on NOCS CAMERA25 and REAL275 [12] training set in a fully-supervised manner to estimate the 6D pose for each unique object in Wild6D. Comparing with most existing work, RePoNet not only has a better generalization ability, but also leverages the in-the-wild data effectively.

B Architecture Details

As describe in our paper, the RePoNet has two parallel branches: *Pose Network* and *Shape Network*. The detailed architecture of each network are illustrated in Fig. 2.

For the Pose Network, given the RGB feature and geometry feature from feature extraction step, we first concatenate them together and feed into a Graph Convolutional Network (GCN) proposed in [6]. We make some modifications on the original GCN: we first construct the deformable kernel based input point clouds as proposed in [6] and then use the concatenated feature as the input to the GCN instead of only point clouds. Here, we use 5 GCN layers in total and concatenate the output from every layer as the final output with dimension of 1792, denoted as $f_{\text{rgbd}} \in \mathbb{R}^{n\times 1792}$, where n is the number of sampling points. Meanwhile, recall the feature of categorical shape prior is obtained via a three-layer PointNet [8] of which dimension is 1024, denoted as $f_{\text{cate}} \in \mathbb{R}^{m\times 1024}$, where m is the number of vertices. Both the number of sampling points and number of vertices are set to 1024. Then we apply a max-pooling operation to f_{cate} along with the point dimension and repeat it back with the number of sampling points for concatenation with f_{rgbd}, denoted as $f_{\text{nocs}} \in \mathbb{R}^{n\times 2186}$. Then, we implement a stack of Multi-Layer Perceptrons(MLPs) with the output channels of $(512, 256, 3)$ as an implicit function with an input point position and the corresponding feature to predict the NOCS coordinates of input point clouds. To further estimate the 6D pose, *i.e.*, rotation ($\mathbf{R}$), translation ($\mathbf{T}$) and scale

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Data split	Bottle	Bowl	Mug	Laptop	Camera	Total
Train set	470	450	450	95	95	1560
Test set	50	50	50	6	6	162
Total	520	500	500	101	101	1722

Table 1: **Number of unique objects for the 5 categories in our Wild6D.**

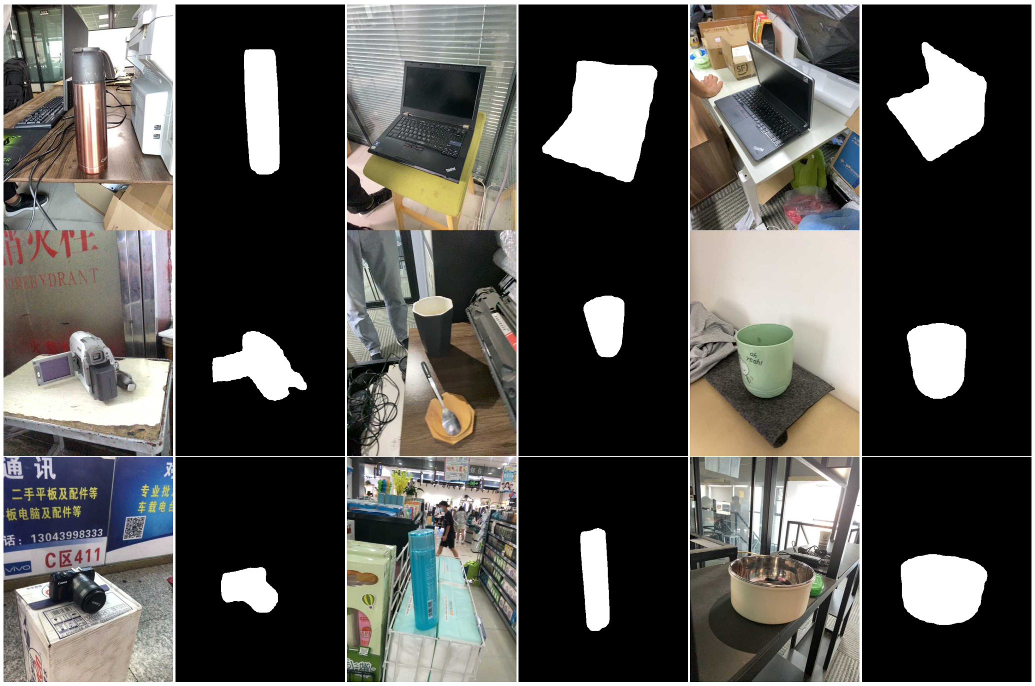

Figure 1: **Segmentation Results on Wild6D.**

(**S**), we directly take concatenate RGBD feature f_{rgbd} with the predicted NOCS coordinates and feed into three parallel convolutional layers whose output channels are $(512, 256, 4)$, $(512, 256, 3)$ and $(512, 256, 1)$. Note that we predict quaternion representation of rotation $\mathbf{R}$.

For the Shape Network, we follow the same step of concatenation obtaining the deformation feature $f_{\text{shape}} \in \mathbb{R}^{m \times 2186}$. Specifically, we first perform the max-pooling over the f_{rgbd} along the point dimension, then concatenate it with per-vertex feature f^{i}_{cate} along the channel dimension. The feature vector after concatenation is used as the input for deformation prediction, denoted as $f_{\text{shape}} \in \mathbb{R}^{m \times 2186}$. And, similarly, three MLPs with the output channels of $(512, 256, 3)$ are used as an implicit function with an mesh vertices position and the corresponding feature to estimate the pre-vertex deformation.

C Implementation Details

Semi-supervised setting. We use the training data of CAMERA25 [12] along with the corresponding annotations and jointly train the model with images of REAL275 [12] or Wild6D without any 6D pose annotations. After cropping the object from the RGBD image, we first resize it to 192×192 and then randomly sample 1,024 points from both color image and depth map. To obtain the categorical shape prior, we choose a CAD model per category from the CAMERA25 training set manually and reduce its number of vertices to 1,024 as well. More configurations of RePoNet have been specified in supplementary materials. We adopt Adam [5] to optimize our model with the initial learning rate of 0.0001. The learning rate is halved every 10 epochs until convergence. We empirically set the balance parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 to $0.2, 2.0, 5.0$ and 0.2, respectively.

Training Details. To extract the RGB feature, we utilize the PSPNetwork [13] with ResNet50 [4] pre-trained on ImageNet [3] as the backbone network. We adopt Adam [5] to optimize our model with the initial learning rate of 0.0001 and halve it every 10 epochs until convergence. The batch

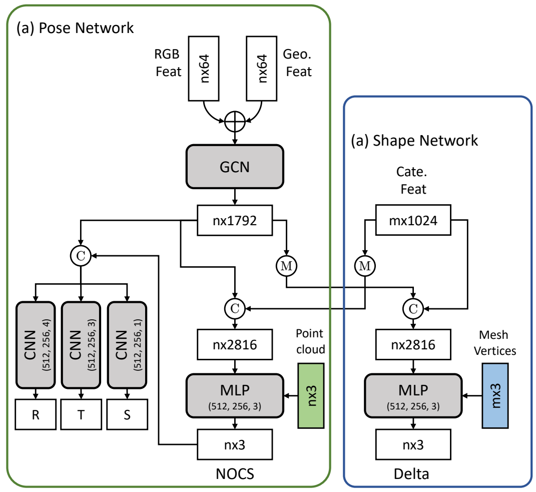

Figure 2: **The architecture details of proposed RePoNet.**

size is set to 32 where the ratio of synthetic data and real-world data is 3 : 1. Besides the disentangle pose loss, NOCS regression loss, shape reconstruction loss and mask loss described in our paper, we have a regularization term on deformed mesh to discourage large deformation: $\mathcal{L}_{\text{reg}} = \frac{1}{N_{\text{v}}} \sum_i^{N_{\text{v}}} \mathrm{m}^i_{\text{delta}}$. By minimizing the predicted deformation, we can preserve more semantic consistency between the categorical shape prior and reconstruction one. Similarly, we have a balance parameter on the regularization term as well and set it to 0.01. Additionally, RePoNet is a category-specific model since RePoNet highly depends on the categorical shape prior and a differentiable rendering module is involved. In other words, we have six models totally for inference on REAL275 [12] and each model only works for a single category.

D More experiments

D.1 Amount of unlabeled real data.

We analyze the effect of using different fractions of unlabeled real data used during semi-supervised learning. We uniformly sample every 10% fraction of collected Wild6D training data for semi-supervised learning and evaluate the performance on REAL275 and Wild6D testing sets. As shown in Fig. 4, with more real data used during training, the object pose estimation performance is getting better.

D.2 Shape reconstruction

To evaluate the shape reconstruction performance, we computed the Chamfer Distance of the reconstructed object mesh with the ground truth one and compared it with other methods. Since the Wild6D does not provide the CAD models, we just conduct this experiment on REAL275 [12]. From Table 2, we observe that the average distance over six categories of our proposed method is much lower than Shape-Prior [11] and SGPA [2] under both fully-supervised setting and semi-supervised setting. However, it is worse than CASS [1], especially on camera category. We believe this is because our method deforms the object shape from the predefined mesh and the shape variance across different cameras is large which may degenerate the performance. Some visualization results on Wild6D samples are shown in Fig. 3.

D.3 Semi-supervised on CO3D

As discussed in Related Work, the recent proposed CO3D [9], although with diverse instances in different categories, is difficult to be used for 6D pose estimation due to the error of depth maps predicted via COLMAP [10]. To validate this point, we compare the RePoNet model trained with CAMERA25 [12] and CO3D with the model trained with CAMERA25 [12] data and Wild6D in Table 3. There is no large performance improvement observed by involving the CO3D data, although

Figure 3: **Qualitative results on Wild6D samples**

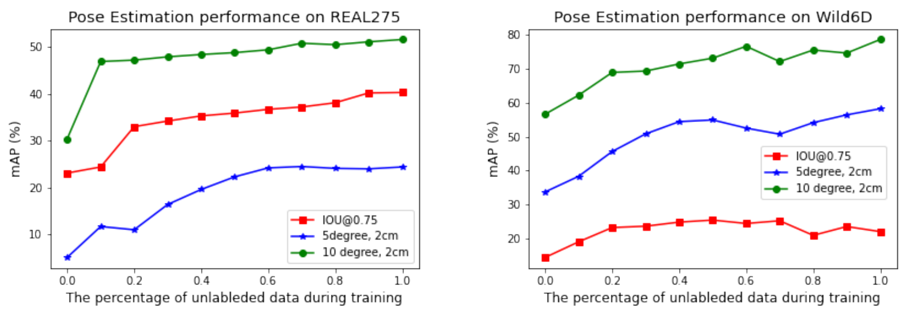

Figure 4: **Ablation study on semi-supervised training with different number of unlabeled real data**. Here, we only show the performance on bottle and evaluate it on both REAL275 and Wild6D dataset.

it also provides hundreds of object-centric real videos. On the other hand, by using the Wild6D data, the pose estimation performance improves a lot. For instance, the improvement on 5 degree, 5cm is 18% versus 3.8%. Hence, our collected Wild6D data with real RGBD images is much more suitable and feasible for 6D pose estimation than existing datasets.

D.4 More visualizations

We show more qualitative results of our proposed RePoNet model on REAL275 [12] and Wild6D in Fig 5, Fig. 6 and Fig. 7. It can be observed that our proposed RePoNet can estimate object pose and size accurately across diverse instances and under d different background scenes.

E Use of existing assets.

We describe the existing assets we used in our paper and the corresponding license of these assets.

CAMERA75&REAL275. Most of experiments are conducted on NOCS dataset collected by [12] which is released on their official website and public to everyone for non-commercial use.

Code. Our code is built upon the Pytorch [7]. And we leverages the code from the released codes from Shape-Prior [11] under the MIT License.

F Personal data and human subjects

We collect a new large-scale RGBD video dataset Wild6D for object pose estimation. The dataset does not include the facial or other identifiable information of humans. We plan to release the collected video dataset if the paper is accepted.

Table 2: **Comparison of shape reconstruction performance.** Numbers show the Chamfer Distance($\times 10^{-3}$) between the estimated shapes and the ground-truth CAD models.

Methods	Bottle	Bowl	Camera	Can	Laptop	Mug	Avg
Shape-Prior [11]	3.44	1.21	8.89	1.56	2.91	1.02	3.17
SGPA [2]]	2.93	0.89	5.51	1.75	1.62	1.12	2.44
CASS [1]	**0.75**	**0.38**	**0.77**	**0.42**	3.73	**0.32**	**1.06**
RePoNet-semi	1.80	0.79	9.50	1.02	2.32	1.24	2.78
RePoNet-sup	1.51	0.76	8.79	1.24	**1.01**	0.94	2.37

Table 3: **Ablation study on CO3D data.** We show the pose estimation performance on bottle when evaluating on REAL275

Training Data CO3D	Training Data Wild6D	$\text{IOU}_{0.75}$	5 degree 2cm	5 degree 5cm
		46.9	15.1	43.1
✓		47.3	17.3	45.8
	✓	**49.1**	**27.3**	**61.1**

References

[1] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020.

[2] 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.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

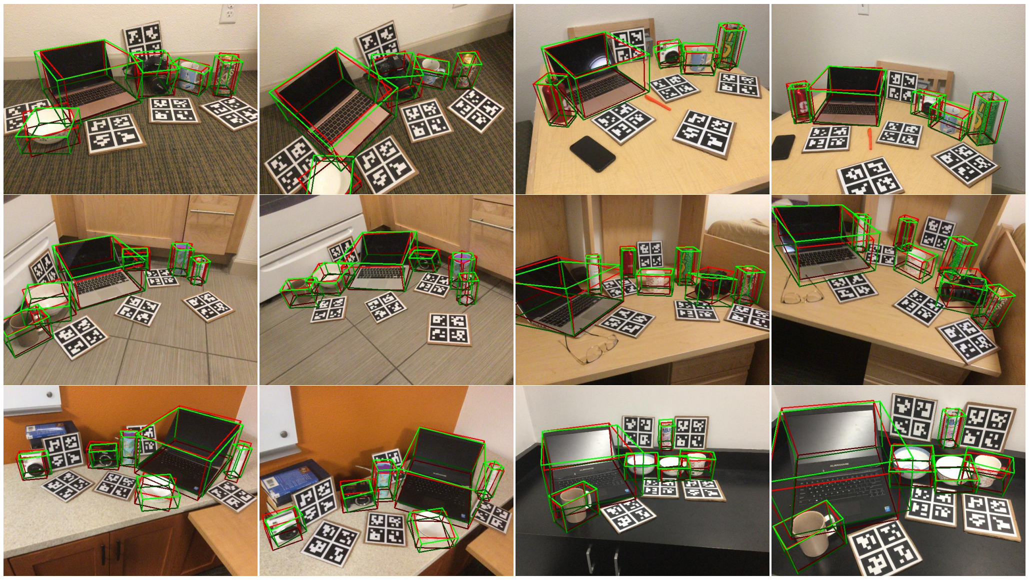

Figure 5: **Visualization Results on REAL275 test set**. Red 3D bounding boxes denote the ground truth, and the green boxes are estimation results via our proposed method.

Figure 6: **Visualization Results on Wild6D test set**. Red 3D bounding boxes denote the ground truth, and the green boxes are estimation results via our proposed method.

[6] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1800–1809, 2020.

[7] 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.

[8] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[9] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021.

[10] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[11] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020.

Figure 7: **Visualization Results on Wild6D test set**. Green 3D bounding boxes denote the ground truth, and the red boxes are estimation results via our proposed method.

[12] 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.

[13] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.