Supplemental Materials
StablePose: Learning 6D Object Poses from Geometrically Stable Patches

Yifei Shi Junwen Huang Xin Xu Yifan Zhang Kai Xu
National University of Defense Technology

1. Details of Stability Analysis

Here, we provide the details of stability analysis of point cloud. Mathematically, given a 3D point set $\mathcal{P} = \{\mathbf{v}_i, \mathbf{n}_i\}$ sampled on the template model surface, we want to find a rigid transformation $[\mathbf{R}|\mathbf{t}]$ which minimizes the following point-to-plane error at all points:

$$\min_{[\mathbf{R},\mathbf{t}]} \sum_i [(\mathbf{R}\mathbf{v}_i + \mathbf{t}) \cdot \mathbf{n}_i]^2, \tag{1}$$

where $\mathbf{R}$ and $\mathbf{t}$ are rotation and translation, respectively.

The rotation $\mathbf{R}$ is nonlinear but can be linearized assuming infinitesimal rotations:

$$\mathbf{R} \approx \begin{pmatrix} 1 & -\gamma & \beta \\ \gamma & 1 & -\alpha \\ -\beta & \alpha & 1 \end{pmatrix}, \tag{2}$$

for Euler angles α, β, and γ around the X, Y, and Z axes, respectively. This reduces the rotation of $\mathbf{v}_i \in \mathbf{V}$ by $\mathbf{R}$ into a displacement of it by a vector $[\mathbf{r} \times \mathbf{v}_i + \mathbf{t}]$, where $\mathbf{r} = (\alpha, \beta, \gamma)$. Substituting this into Equation (1), we therefore aim to find a 6-vector $[\mathbf{r}^T, \mathbf{t}^T]$ that minimizes

$$\min_{[\mathbf{r},\mathbf{t}]} \sum_i [\mathbf{v}_i \cdot \mathbf{n}_i + \mathbf{r} \cdot (\mathbf{v}_i \times \mathbf{n}_i) + \mathbf{t} \cdot \mathbf{n}_i]. \tag{3}$$

This is a linear least-squares problem which amounts to solve a linear system $C\mathbf{x} = 0$ with $\mathbf{x} = [\mathbf{r}^T, \mathbf{t}^T]$. C is a 6×6 covariance matrix of the rigid transformation accumulated over all sample points:

$$C = \sum_i \begin{bmatrix} \mathbf{u}_{ix} \\ \mathbf{u}_{iy} \\ \mathbf{u}_{iz} \\ \mathbf{n}_{ix} \\ \mathbf{n}_{iy} \\ \mathbf{n}_{iz} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{ix} & \mathbf{u}_{iy} & \mathbf{u}_{iz} & \mathbf{n}_{ix} & \mathbf{n}_{iy} & \mathbf{n}_{iz} \end{bmatrix}, \tag{4}$$

where $\mathbf{u} = \mathbf{v} \times \mathbf{n}$. The covariance matrix C encodes the increase of the point-to-plane error when the transformation is moved away from its optimum. The larger the error increase, the less slippable and more stable along that transformation the shape is. On the contrary, if there is a transformation that causes small increase in the error, the shape is unstable w.r.t. the corresponding DoFs.

The stability can then be analyzed by calculating the eigenvalues of C. Let $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_6$ be the eigenvalues of C. The stability is measured as $[1 + e^{0.05\left(\frac{\lambda_6}{\lambda_1} - 200\right)}]^{-1}$, where λ_1 and λ_6 are the smallest and largest eigenvalues of C, respectively. In our method, we select patch group whose stability measure is greater than 0.5 as the geometrically stable patch group.

2. Details of Training Data

Unlike most of the baselines that were trained on both the real images with annotations and the synthetic image rendered by the 3D CAD models, we train StablePose by leveraging the real images only. To be specific, we use the real training images of single objects provided in [2] for `T-LESS` and the real training images of cluttered scenes provided in [1] for `LineMOD-O`. For the training images of single objects in `T-LESS`, we add random occlusion and background to make them more realistic. We experimentally found that training on real data only is sufficient for StablePose. Nevertheless, it is still possible to further boost the performance by utilizing some synthetic data.

3. Visual Ablation of Patch-wise Pose Estimation

To further illustrate the effect of the patch-wise pose estimation component, we qualitatively compare StablePose with *without patch-wise pose estimation* baseline on `T-LESS` in Figure 1. It shows that StablePose produces more accurate results on most shown cases, which demonstrates patch-wise pose estimation indeed provides substantial enhancements for improving object pose accuracy.

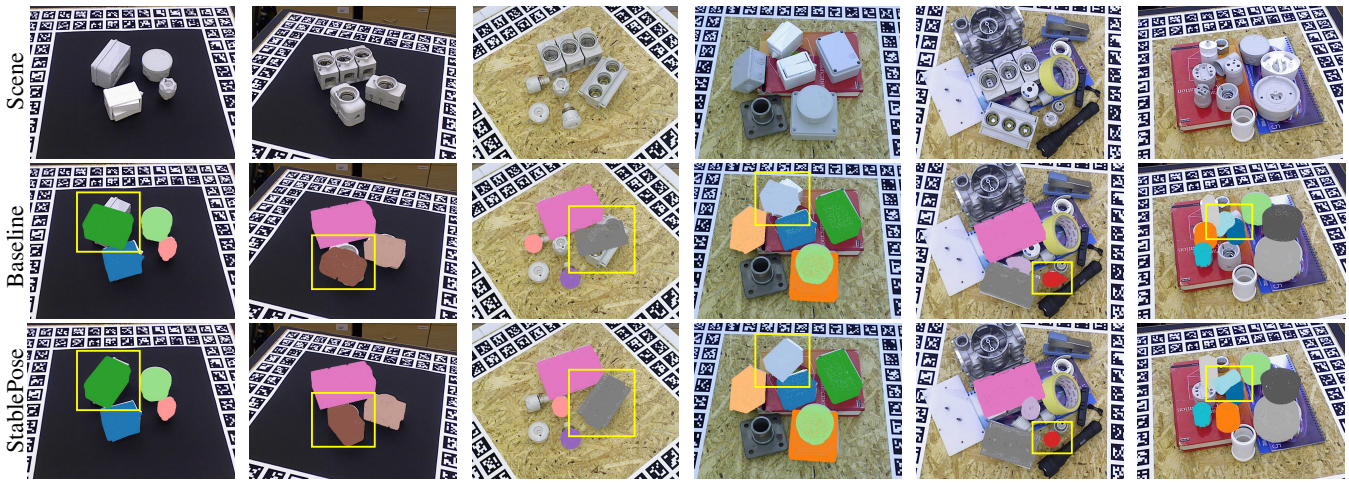

Figure 1: Visual results of 6D object pose estimation by *without patch-wise pose estimation* baseline and StablePose on T-LESS.

4. Quantitative Evaluation of Individual Objects

To further illustrate the strengths and weaknesses of StablePose on different object types, we provide the quantitative results of individual objects of T-LESS in Figure 2. The number above each object is the e_{VSD} (VIVO) value of the corresponding object. The red, green and blue blocks represent high (0.8-1), middle (0.6-0.8) and low (0-0.6) e_{VSD} (VIVO) values, respectively. The results show that StablePose works the best on objects with sufficient number of large planar and/or cylindrical patches. This verifies the effects of our network designs that utilize planar and/or cylindrical patches. The inferior results of composite objects with multiple parts and objects with concave geometry are due to the incapability of extracting repeatable patches from the noisy depth images by the current patch extraction approach.

5. Quantitative Comparison on NOCS-REAL275

We compare StablePose with two baselines on NOCS-REAL275. The baselines are NOCS [4] and 6-pack [3]. Note that while StablePose and NOCS take single RGBD images as input, 6-pack process RGBD videos and estimate the object poses of all frames jointly, which could greatly boost the pose estimation performance. We use the following evaluation metrics proposed in [4, 3]: $5°5cm$, $IoU25$, R_{err} and T_{err}. In Table 1, we report the quantitative results. It shows that StablePose beats all these baselines in two metrics ($5°5cm$, T_{err}) and achieves comparable performances in the others ($IoU25$, R_{err}). This demonstrates the cross-instance generality of StablePose. It also reveals that the generality of category-level pose estimation relies more on geometry instead of appearance.

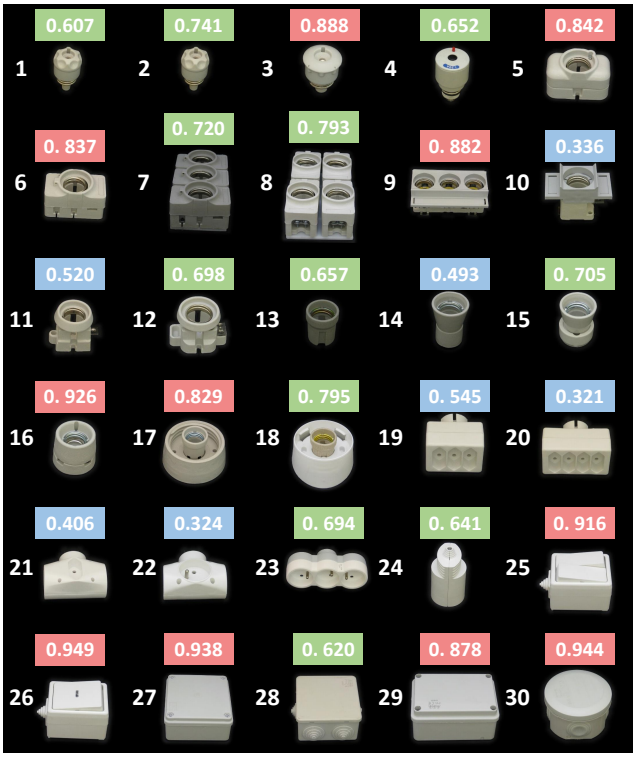

Figure 2: Quantitative results of StablePose for individual objects in T-LESS. The performance is evaluated by e_{VSD} (VIVO). The red, green and blue blocks correspond to high (0.8-1), middle (0.6-0.8) and low (0-0.6) e_{VSD} (VIVO) values, respectively.

6. The ShapeNetPose dataset

6.1. Overview

ShapeNetPose consists of rendered RGBD images of objects from 22 categories. For each category, 80% and

Table 1: Quantitative comparison on NOCS-REAL275.

	Metric	NOCS [4]	6-pack [3]	StablePose
Bottle	$5°5cm$	5.5	24.5	**37.0**
	$IoU25$	48.7	91.1	**94.8**
	R_{err}	25.6	**15.6**	19.4
	T_{err}	14.4	**4.0**	4.5
Bowl	$5°5cm$	62.2	55.0	**76.6**
	$IoU25$	99.6	**100.0**	**100.0**
	R_{err}	4.7	5.2	**4.0**
	T_{err}	1.2	1.7	**1.1**
Camera	$5°5cm$	0.6	**10.1**	4.3
	$IoU25$	**90.6**	87.6	85.5
	R_{err}	**33.8**	35.7	43.9
	T_{err}	**3.1**	5.6	4.5
Can	$5°5cm$	7.1	**22.6**	17.2
	$IoU25$	77.0	**92.6**	90.5
	R_{err}	16.9	**13.9**	20.5
	T_{err}	**4.0**	4.8	4.4
Laptop	$5°5cm$	25.5	63.5	**80.0**
	$IoU25$	94.7	98.1	**99.4**
	R_{err}	8.6	**4.7**	4.8
	T_{err}	2.4	2.5	**2.1**
Mug	$5°5cm$	0.9	**24.1**	17.7
	$IoU25$	82.8	**95.2**	92.9
	R_{err}	31.5	21.3	**19.8**
	T_{err}	4.0	**2.3**	3.8
Overall	$5°5cm$	17.0	33.3	**38.8**
	$IoU25$	82.2	**94.2**	93.9
	R_{err}	20.2	**16.0**	18.7
	T_{err}	4.9	3.5	**3.4**

20% objects are selected for training and testing respectively. All the objects are normalized into a $1m \times 1m \times 1m$ box. Each selected object is rendered from a random viewpoint to generate a RGBD image. Random occlusions are added to each RGBD image. Statistics of ShapeNetPose is listed in Table 2.

6.2. Evaluation Metrics

The results on ShapeNetPose are evaluated by four evaluation metrics: $10°10cm$, $IoU25$, R_{err} and T_{err} as proposed in [4, 3]. Note that, as ShapeNetPose is a synthetic dataset, to facilitate the evaluation, all the objects in ShapeNetPose are normalized into a $1m \times 1m \times 1m$ cube.

7. More Visual Results

Besides the quantitative comparisons, we provide additional visual results by StablePose on T-LESS, LineMOD-O, NOCS-REAL275 and ShapeNetPose in Figure 3. In general, we see that StablePose is capable of handling cases encompassing asymmetric or symmetric objects, objects with occlusion and unseen objects.

Table 2: Statistics of ShapeNetPose.

Category	#Train object	#Test object
airplane	500	100
ashcan	200	40
bag	50	10
bathtub	500	100
bed	100	20
bench	500	100
bottle	400	80
bus	500	100
camera	80	16
can	80	16
car	500	100
chair	500	100
display	500	100
earphone	50	10
guitar	500	100
helmet	100	20
lamp	500	100
laptop	300	60
pot	500	100
skateboard	100	20
sofa	500	100
table	500	100

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014. 1

[2] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 1

[3] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020. 2, 3

[4] 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 Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 2, 3

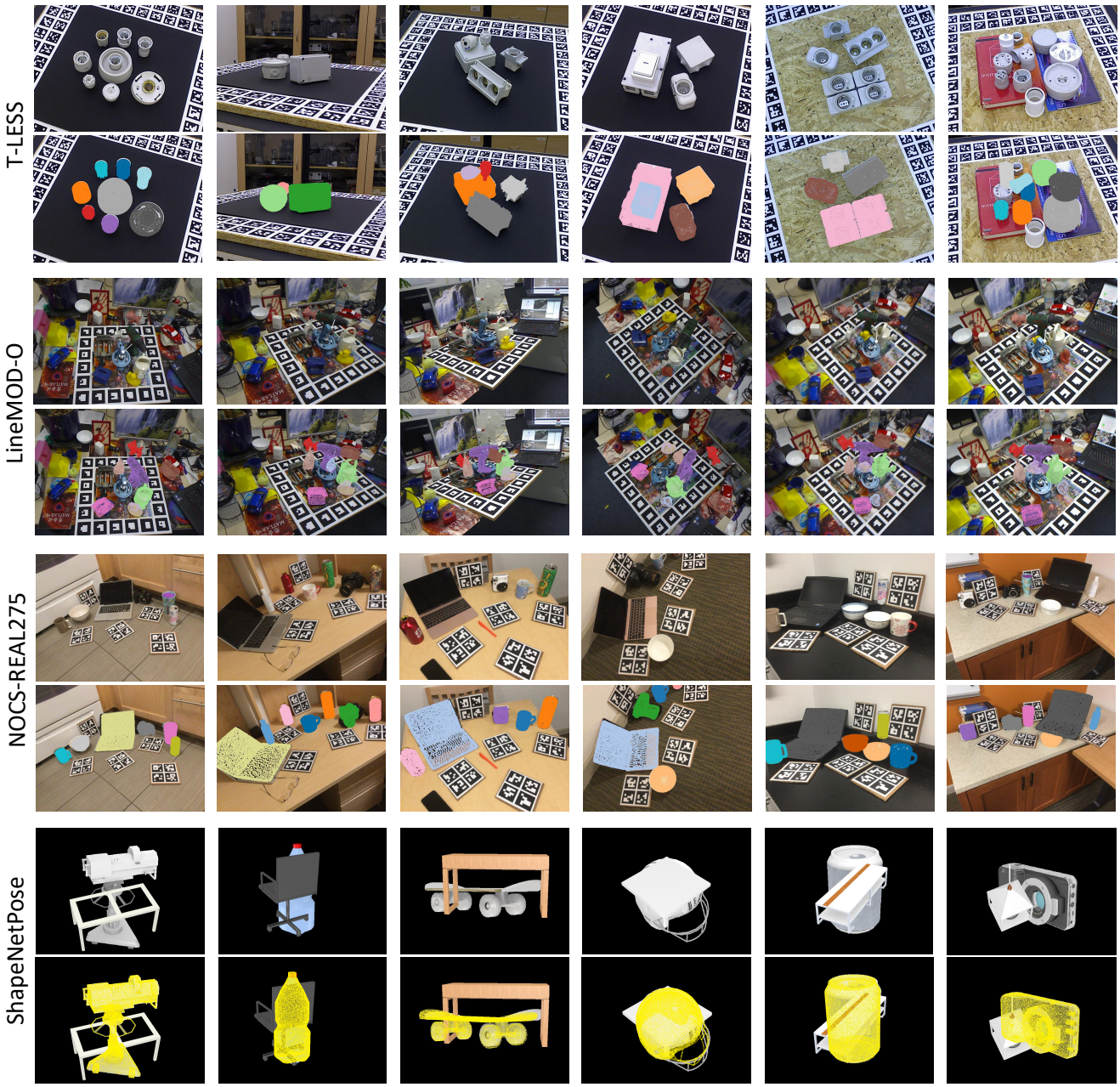

Figure 3: More visual results of 6D object pose estimation by StablePose on `T-LESS`, `LineMOD-O`, `NOCS-REAL275` and `ShapeNetPose`.