

# Appendix— *UnPose*: Uncertainty-Guided Diffusion Priors for Zero-Shot Pose Estimation

Anonymous Author(s)

Affiliation

Address

email

## 1 A Additional Qualitative Results

2 **Scene-level Reconstructions.** To demonstrate the generalization capabilities of *UnPose*, we test  
3 its performance beyond tabletop objects (YCB-V, LM-O) on large-scale furniture objects from the  
4 ScanNet [1]. As shown in Fig. 1, *UnPose* reconstruct multiple diverse objects in a scene, including  
5 four chairs, one sofa, and one table, all from the same pipeline. We showcase the reconstruction of  
6 one object (chair) in the scene, and show it from an input camera viewpoint (Front View) and two  
7 novel views (Back View 1 and 2). The smaller floater images are real observations and diffusion  
8 frames provided to backend optimization, offering context on the input data.

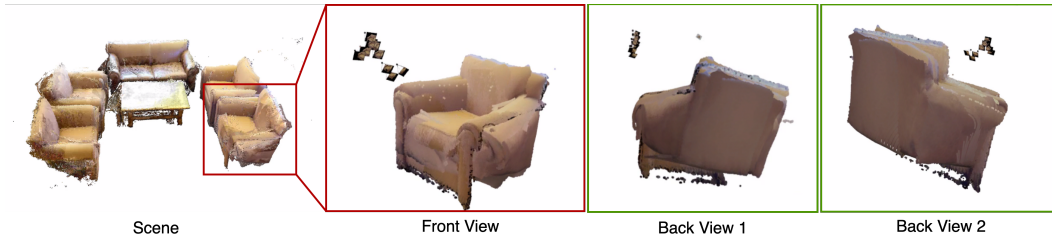


Figure 1: Visualization of a scene-level reconstruction on a ScanNet scene [1] by *UnPose*. We show the reconstruction of one of the chairs from one input camera viewpoint (Front View) as well as diffused novel views (Back View 1 and 2). Real and diffusion images of the scene used in backend optimization are shown as floaters.

9 Fig. 2 visualizes the pixel-level uncertainty estimations associated with the novel views generated  
10 by the diffusion model [2] used in our pipeline. For a better understanding, we also show the ground  
11 truth renderings of the object alongside each diffused view. It can be seen that the diffusion model  
12 exhibits higher variance (greater uncertainties) for novel views that significantly deviates from the  
13 input. This implies the model is less confident when synthesizing unseen regions of the object.

14 We further provide qualitative comparisons with SOTA 6D pose estimation methods [3, 4, 5], includ-  
15 ing GigaPose [3], SAM-6D [4], and FoundationPose [5] on standard 6D pose estimation benchmark  
16 datasets: YCB-Video [6] and LM-O [7]. Fig. 3 clearly shows that compared to prior methods where  
17 the estimated translation and rotation show high error, our proposed method, *UnPose*, accurately  
18 estimates the 6DOF of the object.

## 19 B Additional Quantitative Results

20 In this section, we present additional quantitative comparison details on YCB-Video [6] and LM-  
21 O [7] datasets. We report object reconstruction results on YCB-V [6] dataset in Tab. 1, evaluating  
22 performance with 1, 8 and 16 images. Our proposed method consistently achieves lower reconstruc-  
23 tion error (measured as Chamfer Distance) and is significantly faster than prior methods [8, 9, 2].

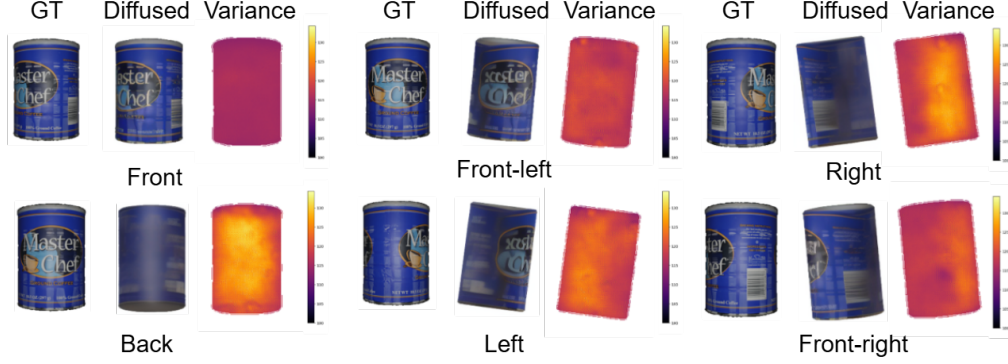


Figure 2: Visualization of diffused views, the associated pixel-level uncertainty estimates, and the corresponding ground-truth rendering perspectives. The diffused images show larger variance at unseen angles.

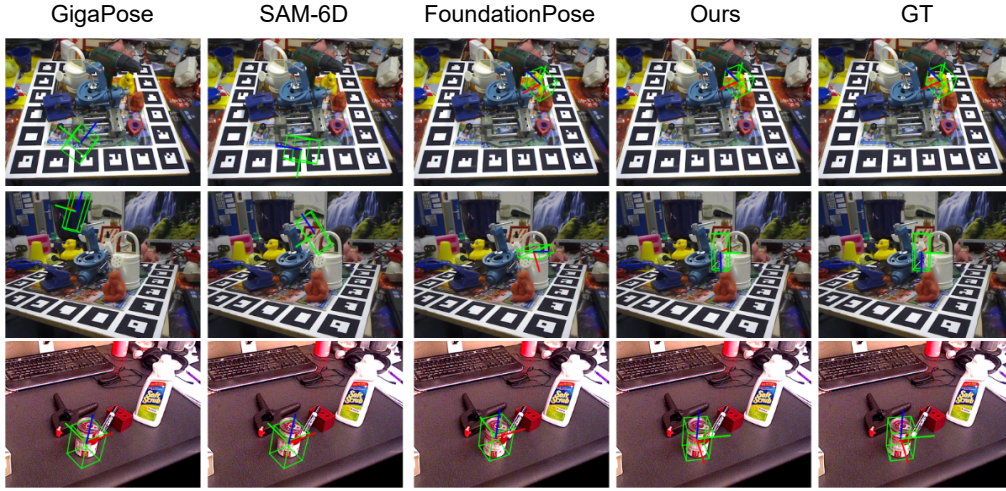


Figure 3: Qualitative results of *UnPose* compared to prior pose estimation methods on YCB-Video [6] and LM-O [7] datasets when 8 reference views are used.

We tabulate the 6D pose estimation results on YCB-Video [6] and LM-O [7] in Tabs. 2 and 3 respectively. Following prior works [5, 10, 11], we evaluate the performance on pose estimation using ADD and ADD-S. Furthermore, for each method we report the results when 1, 8 and 16 images are used for reconstruction [3, 5] or as reference views [4]. It can be seen in Tabs. 2 and 3 that our method consistently outperforms prior SOTA methods (visible as gradual increase in the brightness of the gradient) in all settings *i.e.*, both metric and the number of images.

Method	Metric	$ycbv_2$			$ycbv_6$			$ycbv_7$			$ycbv_8$			$ycbv_9$			$ycbv_{10}$		
		1	8	16	1	8	16	1	8	16	1	8	16	1	8	16	1	8	16
Wonder3d [2]	CD	0.23	0.00	0.00	0.17	0.00	0.00	0.12	0.00	0.00	0.13	0.00	0.00	0.07	0.00	0.00	0.07	0.00	0.00
	Time (s)	128.20	0.00	0.00	130.63	0.00	0.00	132.22	0.00	0.00	125.30	0.00	0.00	126.40	0.00	0.00	130.20	0.00	0.00
BundleSDF [9]	CD	0.03	0.02	0.02	0.07	0.04	0.02	0.03	0.02	0.01	0.03	0.02	0.01	0.05	0.02	0.01	0.00	0.00	0.00
	Time (s)	106.35	155.27	172.96	101.74	112.90	126.31	104.23	118.74	131.49	109.20	114.46	130.76	104.97	129.13	147.59	102.11	115.84	132.76
GOM [8]	CD	0.06	0.12	0.08	0.15	0.12	0.10	0.06	0.07	0.08	0.04	0.17	0.09	0.04	0.04	0.05	0.15	0.27	0.17
	Time (s)	28.88	53.64	121.19	28.63	54.92	128.65	28.38	53.53	120.24	28.39	52.47	116.98	27.97	53.11	115.53	28.27	52.58	117.18
Ours	CD	0.02	0.00	0.00	0.03	0.02	0.02	0.01	0.01	0.00	0.02	0.01	0.01	0.01	0.01	0.00	0.02	0.00	0.00
	Time (s)	13.20	26.20	41.20	13.80	27.20	41.70	12.90	26.30	43.22	13.50	25.30	40.25	14.10	22.21	41.30	11.30	23.20	42.10

Table 1: Quantitative comparison for reconstruction with 1, 8 and 16 images using *UnPose* in terms of accuracy (reported using chamfer distance (CD) as  $10^{-3}$ ) and time (sec) on objects from YCB-V [6]. For both CD and time, brighter gradients denotes better performance.

Method	# img	Metric	Mean	005_tomato_soup_can	006_mustard_bottle	007_tuna_fish_can	009_gelatin_box	010_potted_meat_can	011_banana
GigaPose [3]	1	ADD	13.58	8.20	4.15	1.77	45.92	4.72	16.74
		ADD-S	24.44	26.23	10.66	5.42	66.69	14.17	23.48
	8	ADD	21.48	17.74	14.03	3.33	75.00	6.62	12.16
		ADD-S	39.52	58.06	28.87	14.35	79.62	22.47	33.78
	16	ADD	60.58	88.80	62.38	13.67	93.75	62.35	42.56
		ADD-S	86.8	96.82	100.00	49.33	100.00	92.35	82.33
FoundationPose [5]	1	ADD	30.34	1.69	40.00	7.69	76.00	16.67	40.00
		ADD-S	74.58	42.01	86.90	71.10	100.00	72.10	75.40
	8	ADD	45.57	44.07	55.00	41.00	80.00	13.33	40.00
		ADD-S	90.17	76.20	87.33	91.20	100.00	89.10	97.20
	16	ADD	71.2	62.10	66.00	72.30	100.00	60.60	66.20
		ADD-S	95.56	98.37	95.00	100.00	100.00	80.00	100.00
SAM6D [4]	1	ADD	19.93	11.66	37.50	3.75	31.60	10.00	25.10
		ADD-S	43.04	13.33	86.83	16.76	52.10	50.00	39.20
	8	ADD	34.63	16.67	44.67	6.87	76.25	38.33	25.00
		ADD-S	63.86	36.67	82.33	31.00	96.70	65.17	71.28
	16	ADD	72.06	60.27	80.20	77.10	100.00	56.36	58.44
		ADD-S	97.02	98.42	100.00	100.00	100.00	83.73	100.00
Ours	1	ADD	47.92	20.34	55.00	58.20	80.00	44.00	50.00
		ADD-S	82.72	82.12	90.00	69.23	100.00	80.00	75.00
	8	ADD	61.2	56.60	70.00	53.60	87.00	40.00	60.00
		ADD-S	91.22	88.22	100.00	89.13	100.00	90.00	80.00
	16	ADD	74.39	64.98	80.00	66.67	91.00	76.67	67.00
		ADD-S	96.65	89.83	100.00	94.87	100.00	100.00	95.20

Table 2: Quantitative performance of *UnPose* compared to SOTA baselines on YCB-Video [6] for 6D Pose estimation. For both ADD and ADD-S metrics, we show higher values with brighter gradient and vice-versa.

Method	# imgs	Metric	Mean	bath duck	cat toy	hole puncher	power drill	water can
Gigapose [3]	1	ADD	19.25	1.07	1.15	2.53	48.65	42.86
		ADD-S	34.77	6.82	6.90	8.63	83.78	67.72
	8	ADD	33.46	15.56	4.87	10.84	66.67	69.39
		ADD-S	54.07	33.37	12.43	50.21	90.90	83.45
	16	ADD	67.29	63.82	57.47	29.36	89.47	96.36
		ADD-S	87.53	97.28	97.70	47.37	97.37	97.96
FoundationPose [5]	1	ADD	34.00	8.70	26.12	22.00	82.23	30.95
		ADD-S	71.28	34.78	78.10	70.10	90.13	83.33
	8	ADD	40.79	19.57	22.78	18.26	88.23	55.14
		ADD-S	82.60	78.27	86.83	66.26	96.33	85.33
	16	ADD	84.29	76.26	78.30	94.26	81.20	91.46
		ADD-S	97.08	95.65	94.13	98.12	100.00	97.52
SAM6D [4]	1	ADD	29.16	6.13	16.49	33.23	50.47	39.50
		ADD-S	61.69	14.55	60.82	94.10	67.20	71.81
	8	ADD	37.00	11.35	31.96	42.13	52.12	47.47
		ADD-S	70.35	33.67	64.95	95.10	76.25	81.81
	16	ADD	82.65	69.21	69.16	88.36	96.12	90.40
		ADD-S	93.00	90.22	76.29	99.93	99.59	98.99
Ours	1	ADD	50.11	40.30	34.22	40.10	83.63	52.31
		ADD-S	84.20	66.67	85.23	88.30	92.10	88.70
	8	ADD	67.27	52.21	52.10	71.10	86.67	74.27
		ADD-S	91.58	87.70	90.12	90.20	97.30	92.62
	16	ADD	85.00	77.21	81.10	88.20	90.21	88.30
		ADD-S	98.19	97.34	95.33	98.30	100.00	100.00

Table 3: Quantitative performance of *UnPose* compared to SOTA baselines on LM-O [7] for 6D Pose estimation. For both ADD and ADD-S metrics, we show higher values with brighter gradient and vice-versa.

## 30 C Deployments on a Real-world Robotic Platform

31 We further demonstrate the effectiveness and the deployment capability of *UnPose* in a real-world  
32 robotic manipulation pipeline using a PiPER robot arm. In this setup, *UnPose* estimates the complete  
33 3D geometry and 6D pose of a target object from RGB-D stream inputs captured by a wrist-mounted  
34 Intel RealSense D435 camera. This geometric and pose information is then passed to a subsequent  
35 grasp planning module, AnyGrasp [12], to determine a suitable grasp pose for the robot. Fig. 4  
36 provides a visual snapshot of this system in action, showcasing *UnPose*'s reconstruction and pose



Figure 4: Visualization of *UnPose* for a real-world robotic manipulation task where it estimates the complete geometry and 6D pose of the cup.

estimation for a target mug on PiPER. Further details, including the dynamic operation of the system capturing multiple views, the resulting reconstruction, and the robot successfully grasping the object, are presented in the accompanying video.

## References

- [1] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1
- [2] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 1, 2
- [3] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024. 1, 2, 3



- [4] J. Lin, L. Liu, D. Lu, and K. Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 1, 2, 3
- [5] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 1, 2, 3
- [6] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 1, 2, 3
- [7] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014. 1, 2, 3
- [8] Z. Liao, B. Xu, and S. L. Waslander. Toward general object-level mapping from sparse views with 3d diffusion priors. In *8th Annual Conference on Robot Learning*, 2024. 1, 2
- [9] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023. 1, 2
- [10] Y. Liu, Z. Jiang, B. Xu, G. Wu, Y. Ren, T. Cao, B. Liu, R. H. Yang, A. Rasouli, and J. Shan. Hippo: Harnessing image-to-3d priors for model-free zero-shot 6d pose estimation, 2025. URL <https://arxiv.org/abs/2502.10606>. 2
- [11] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon. Any6d: Model-free 6d pose estimation of novel objects. *CVPR*, 2025. 2
- [12] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics (T-RO)*, 2023. 3