DVMNet: Computing Relative Pose for Unseen Objects Beyond Hypotheses

Supplementary Material

6D Relative Object Pose Estimation

As introduced in the main paper, we are interested in extending the presented DVMNet to 6D relative object pose estimation in our future work. Notably, we followed the setting proposed in RelPose, RelPose++, and 3DAHV, assuming the availability of known object positions. Such information provides a strong prior, making the relation translation estimation less challenging. In this scenario, the primary challenge is the detection of an unseen object present in the query image. We achieve the zero-shot object detection by utilizing the detection module presented in Gen6D [21]. Specifically, the detection module predicts the 2D object center and 2D object scale, from which we compute the 3D object translation, following the implementation of Gen6D. The 3D object rotation is estimated by employing our DVMNet. It is worth noting that our goal is to achieve 6D object pose estimation in the single-reference setup, while Gen6D's detection module utilizes dense-view references by default. Therefore, in our experiments, we feed a single reference to the detection module as the template.

We illustrate some preliminary results in Fig. 8 and Fig. 9, visualizing the 6D object pose in the query image as a 3D object bounding box. As shown in Fig. 8, the combination of Gen6D's object detector and our DVMNet achieves promising results in some cases. However, in some scenarios shown in Fig. 9, Gen6D's detector fails to effectively detect the unseen object depicted in the query image. Notably, the zero-shot object detection becomes more challenging in the single-reference setting due to the large pose difference between the query object and the reference. In this context, we plan to investigate the applicability of other zero-shot object detectors such as SAM [17] in future work.

Extension to Sparse-View References

As introduced in the literature [18, 27, 47], some downstream tasks such as 3D reconstruction often rely on sparse-view references. Intuitively, our DVMNet can be seamlessly integrated into these tasks, as it operates effectively with just a single reference. Therefore, we develop an experiment on the CO3D dataset, evaluating DVMNet with varying numbers of reference images, ranging from 1 to 7.

Specifically, given an unseen object during testing, we randomly sample n images. These images are then fed into the presented DVMNet, with one image designated as the query and the remaining ones as references. The object pose in the query image is simply derived from the resulting $n-1$ relative object poses as

$$\mathbf{R}_q = m(\frac{1}{n-1}\sum_{i=1}^{n-1} q(\Delta\mathbf{R}_i\mathbf{R}_r^i)), \tag{18}$$

where $\Delta\mathbf{R}_i$ and $\mathbf{R}_r^i$ denote the i-th relative object pose and reference pose, respectively, $q(\cdot)$ represents a function that converts a rotation matrix to the 6D continuous representation [53], and $m(\cdot)$ indicates the conversion from the 6D continuous representation to a rotation matrix. We

Figure 8. **Visualization of 6D pose estimation for unseen objects on LINEMOD [15].** The ground-truth 6D object pose and the predicted pose in the query image are depicted as green and blue 3D bounding boxes, respectively.

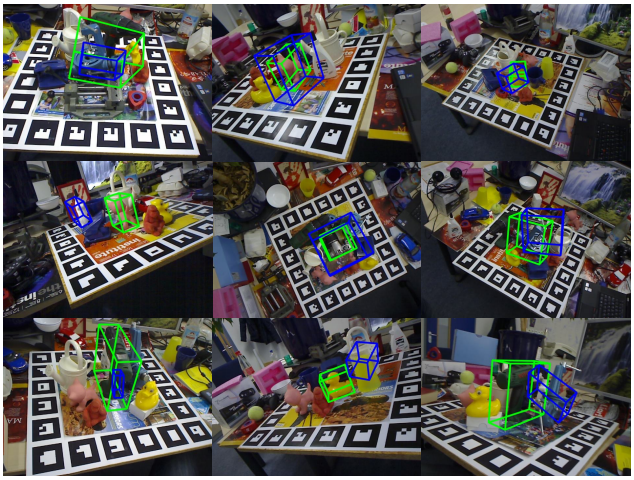

Figure 9. **Failure cases of 6D pose estimation for unseen objects on LINEMOD [15].**

# References	1	2	3	4	5	6	7
SuperGlue	71.47	73.08	68.72	65.90	64.54	63.24	61.74
3DAHV	28.44	29.29	28.20	27.21	26.40	24.85	24.76
DVMNet	**19.95**	**18.38**	**16.79**	**16.21**	**15.73**	**14.99**	**14.93**

Table 4. **Extension to sparse-view references.** The experiment is conducted on CO3D [25] with the number of reference images varying from 1 to 7. The metric employed is the angular error between the computed query object pose and the ground truth.

Method	SuperGlue	3DAHV	DVMNet
Angular Error ↓	73.72	51.49	49.02

Table 5. **Experimental results on the LINEMOD-O [5] dataset.** The mean angular errors are reported.

Method	Acc @ 30° (%) ↑
RelPose	64.2
RelPose++	77.0
3DAHV	83.5
PoseDiffusion	81.8
DVMNet	**84.7**

Table 6. **Additional experimental results on CO3D.** Acc @ 30° is employed as a metric.

also evaluate the representative image-matching (SuperGlue) and hypothesis-based (3DAHV) approaches in the sparse-view scenario. We ensure a fair comparison by utilizing the same strategy of query object pose estimation for these methods.

We report the resulting angular errors in Table 4. It is evident that (i) the angular error of our DVMNet decreases as more reference images are involved, and (ii) DVMNet consistently yields the smallest angular error. This observation demonstrates the promising compatibility of our approach with sparse-view reference images.

Robustness to Occlusions

Given that object pose estimation is often challenged by occlusions, we assess the robustness in scenarios involving occlusions by conducting an experiment on the LINEMOD-O [5] dataset. The testing data comprises three unseen objects, i.e., cat, driller, and duck. We report the mean angular errors of the evaluated methods in Table 5. Our DVMNet outperforms both the image-matching method, SuperGlue, and the hypothesis-based method, 3DAHV, showcasing better robustness against occlusions.

Additional Results on CO3D

As listed in Tabel 6, we report more results on CO3D, using Acc @ 30° as a metric. Note that PoseDiffusion [37] takes multiple views (> 2) as input by default, while all the other evaluated methods employ two views. For a fair comparison, we evaluate PoseDiffusion on CO3D using two views. Moreover, The pose parameters are iteratively updated 100 times during the denoising process in PoseDiffusion, making the method time-consuming compared with our single forward pass mechanism.

References

[1] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7163–7172, 2019. 5

[2] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997. 1

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 4

[5] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Proceedings of the European Conference on Computer Vision*, pages 536–551. Springer, 2014. 2

[6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[7] 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

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 5

[9] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 2, 7, 8

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[11] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1

[12] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 516–532. Springer, 2022. 3, 6, 7, 8

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2, 3, 4

[14] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022. 1, 2

[15] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision*, pages 548–562. Springer, 2012. 2, 7, 8, 1

[16] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2020. 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3, 1

[18] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 1, 2, 3, 6, 7, 8

[19] 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 *Proceedings of the European Conference on Computer Vision*, pages 19–34. Springer, 2022. 2, 5

[20] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. *arXiv preprint arXiv:2303.11328*, 2023. 7

[21] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[23] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE Transactions on Visualization and Computer Graphics*, 22(12):2633–2651, 2015. 1

[24] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 2

[25] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 6, 7

[26] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature

matching with graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4938–4947, 2020. 2, 3, 6, 7, 8

[27] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1

[28] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, page 435. Seattle, WA, 2009. 5

[29] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 1, 2

[30] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 2

[31] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 2, 3, 6, 7, 8

[32] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 1, 2

[33] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *Conference on Robot Learning*, 2018. 1

[34] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 2

[35] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2

[36] 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

[37] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2

[38] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3523–3532, 2019. 4, 5

[39] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11443, 2023. 3

[40] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Proceedings of the European Conference on Computer Vision*, pages 230–247. Springer, 2020. 3

[41] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023. 3

[42] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3109–3118, 2015. 2, 3, 6

[43] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2

[44] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3D objects. In *British Machine Vision Conference (BMVC)*, 2019. 3

[45] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2666–2674, 2018. 3

[46] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5845–5854, 2019. 3

[47] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 592–611. Springer, 2022. 1, 2, 3, 6, 7, 8

[48] Chen Zhao, Zhiguo Cao, Chi Li, Xin Li, and Jiaqi Yang. Nm-net: Mining reliable neighbors for robust feature correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 215–224, 2019. 3

[49] Chen Zhao, Yixiao Ge, Feng Zhu, Rui Zhao, Hongsheng Li, and Mathieu Salzmann. Progressive correspondence pruning by consensus learning. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 6464–6473, 2021. 3

[50] Chen Zhao, Yinlin Hu, and Mathieu Salzmann. Locposenet: Robust location prior for unseen object pose estimation. *arXiv preprint arXiv:2211.16290v2*, 2022. 3

[51] Chen Zhao, Yinlin Hu, and Mathieu Salzmann. Fusing local similarities for retrieval-based 3d orientation estimation of unseen objects. In *Proceedings of the European Conference on Computer Vision*, pages 106–122. Springer, 2022. 2, 3, 6

[52] Chen Zhao, Tong Zhang, and Mathieu Salzmann. 3d-aware hypothesis & verification for generalizable relative object pose estimation. *arXiv preprint arXiv:2310.03534*, 2023. 1, 2, 3, 6, 7, 8

[53] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3, 6, 1