FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects

Supplementary Material

5.1. Performance on BOP Leaderboard

Fig. 8 presents our results on the BOP challenge of "6D localization of unseen objects".[2] Our FoundationPose is #1 on the leaderboard. This corresponds to one of the four tasks considered in this work: model-based pose estimation for novel objects. We use the 2D detection from CNOS [45], which is the default provided by the BOP challenge.

5.2. Implementation Details

During training, we first pretrain the neural object field, pose refine network, and pose ranking network separately. We then perform end-to-end fine-tuning for another 5 epochs, while freezing the weights of the neural object field to only provide rendering on-the-fly. The whole training process is conducted over synthetic data which takes about a week on 4 NVIDIA V100 GPUs. At test time, the model is directly applied to the real world data and runs on one NVIDIA RTX 3090 GPU. Under the few-shot setup, rendering is obtained from the neural object field. Under the model-based setup, rendering is obtained via conventional graphics pipeline [32].

Neural Object Field. We normalize the object into the neural volume bound of $[-1, 1]$. The geometry network Ω consists of two-layer MLP with hidden dimension 64 and ReLU activation except for the last layer. The intermediate geometric feature $f_{\Omega(\cdot)}$ has dimension 16. The appearance network Φ consists of three-layer MLP with hidden dimension 64 and ReLU activation except for the last layer, where we apply sigmoid activation to map the color prediction to $[0, 1]$. We implement the multi-resolution hash encoding [43] in CUDA and simplify to 4 levels, with number of feature vectors from 16 to 128. Each level's feature dimension is set to 2. The hash table size is set to 2^{22}. In each iteration the ray batch size is 2048. The truncation distance λ is set to 1 cm. In the training loss, $w_e = 1, w_s = 1000, w_c = 100$. Training takes about 2k steps which is often within seconds.

[2] https://bop.felk.cvut.cz/leaderboards/pose-estimation-unseen-bop23/core-datasets/

BOP: Benchmark for 6D Object Pose Estimation

HOME CHALLENGES DATASETS LEADERBOARDS SUBMIT RESULTS Sign in

Tasks on seen objects: 6D localization of seen objects 2D detection of seen objects 2D segmentation of seen objects
Tasks on unseen objects: 6D localization of unseen objects 2D detection of unseen objects 2D segmentation of unseen objects

Datasets: Core datasets LM LM-O T-LESS ITODD HB HOPE YCB-V RU-APC IC-BIN IC-MI TUD-L TYO-L

6D localization of unseen objects – Core datasets

This leaderbord shows the overall ranking for Task 4 on the core datasets (LM-O, T-LESS, TUD-L, IC-BIN, ITODD, HB, YCB-V). For each method, the date of the latest considered submission is reported. If more submissions of a method are available for a dataset, the submission with the highest AR_{Core} score is considered. The reported time is the average image processing time averaged over the core datasets.

Show 50 entries Search:

	Date (UTC)	Method	Test image	AR_{Core}	AR_{LM-O}	AR_{T-LESS}	AR_{TUD-L}	AR_{IC-BIN}	AR_{ITODD}	AR_{HB}	AR_{YCB-V}	Time (s)
1	2023-11-19	FoundationPose	RGB-D	0.726	0.733	0.617	0.906	0.528	0.609	0.809	0.882	
2	2023-11-17	PoMZ	RGB-D	0.692	0.684	0.521	0.945	0.503	0.559	0.791	0.840	
3	2023-11-18	SAM6D	RGB-D	0.683	0.687	0.498	0.874	0.561	0.577	0.754	0.828	1.950
4	2023-09-28	GenFlow-MultiHypo16	RGB-D	0.674	0.635	0.521	0.862	0.534	0.554	0.779	0.833	34.578
5	2023-09-25	GenFlow-MultiHypo	RGB-D	0.662	0.622	0.509	0.849	0.524	0.544	0.770	0.818	21.457
6	2023-09-27	Megapose-CNOS_fastSAM+Multihyp_Te...	RGB-D	0.628	0.626	0.487	0.851	0.467	0.468	0.730	0.764	141.965
7	2023-09-27	Megapose-CNOS_fastSAM+Multihyp_Te...	RGB-D	0.623	0.620	0.485	0.846	0.462	0.460	0.725	0.764	116.564
8	2023-09-28	SAM6D-BOP2023-CNOSmask	RGB-D	0.616	0.648	0.483	0.794	0.504	0.351	0.727	0.804	3.872
9	2023-10-18	GenFlow-MultiHypo16-RGB	RGB	0.576	0.572	0.528	0.688	0.458	0.398	0.746	0.642	40.528

Figure 8. Screenshot on BOP leaderboard. At the time of submission, our approach outperforms the previous best method "PoMZ" (not yet published) by a considerable margin of 0.03 on AR_{Core}, setting a new benchmark record on the leaderboard.

Pose Hypothesis Generation. For global pose initialization, $N_s = 42, N_i = 12$. To train the refinement network, the pose is randomly perturbed by adding translation noise under the magnitude of $0.02m, 0.02m, 0.05m$ for XYZ axis respectively and rotation under the magnitude of $20°$, where the direction is randomized. Both the rendering and input observation are cropped based on the perturbed pose and resized into 160×160 before sending to the network. In the training loss (Eq. 10), w_1 and w_2 are both set to 1. The individual training stage takes 50 epochs. The refinement iteration is set to 1 for training efficiency, At test time, it is set to 5 for pose estimation and 1 for tracking. The complete network architecture of the pose refinement module can be found in the main paper (Fig. 2), where the network architecture used for image feature embedding is illustrated in Fig. 9. In the transformer encoder, the embedding dimension is 512, number of heads is 4, feed-forward dimension is 512.

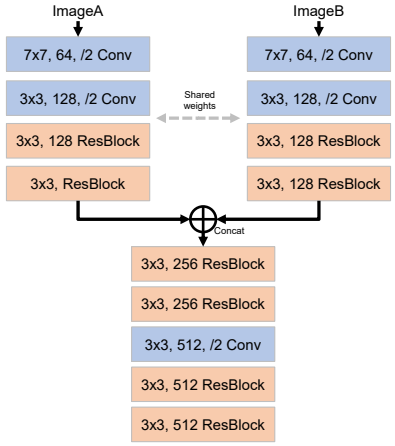

Figure 9. Network architecture for image feature embedding used in pose refinement and selection networks. The ResBlock is from ResNet-34 [16].

Pose Selection. The individual training for the selection network takes 25 epochs, where we perform the similar pose perturbation to refinement network, and the number of pose hypotheses $K = 5$. During the end-to-end fine-tuning, the pose hypotheses come from the output of the refinement network. In the training loss (Eq. 11), α is set to 0.1. The valid positive sample's rotation threshold d is set to $10°$. The complete network architecture of the pose refinement module can be found in the main paper (Fig. 2), where the network architecture used for image feature embedding is illustrated in Fig. 9. When performing the two-level hierarchical comparison, we use the same architecture for both self-attention modules. Concretely, the embedding dimension is 512, number of heads is 4, feed-forward dimension is 512.

Pose Tracking. Our framework can be trivially adapted to the pose tracking task while leveraging temporal cues. To do so, at each timestamp, we send the cropped current frame and the rendering using the previous pose to the pose refinement module. The refined pose becomes the current pose output. This operation repeats along the video sequence. The first frame's pose can be initialized by our pose estimation mode.

Synthetic Data. Objaverse assets vary extremely in the object size and mesh complexity. Therefore, we further normalize the objects and remove the disconnected components automatically based on the mesh edge connectivity graph, to make the objects suitable for learning pose estimation. To create each scene, we randomly sampled 70 to 90 objects and dropped them onto a platform with invisible walls until the object velocities were smaller than a threshold. We randomly scaled the objects from 5 to 30 cm and sampled the size of the platform between 1 to 1.5 meter. The LLM-aided texture augmentation is applied to each object from Objaverse [6] with 3 to 5 different seeds for various styles. To produce diverse and photorealistic images, we randomly created 0 to 5 lights with varied size, color, intensity, temperature and exposure, and $N_c = 2$ cameras on a hemisphere with radius ranging from 0.2 to 3.0 meter above the platform. We also randomize the material properties, including metallicness and reflection, and textures of the objects and the platform. For the environment, we created a dome light with a random orientation and sampled the background from 662 HDR images obtained from Poly Haven [15]. In addition to RGBD rendering, we also store the corresponding object segmentation, camera parameters and the object poses similar to [25, 31]. In total, our dataset has about 600K scenes and 1.2M images. The dataset will be released on the project page upon acceptance.

Creating Reference Images. In the model-free few-shot setup, similar to [21], on YCB-Video and LINEMOD datasets, we select a subset of reference images $\mathbb{S}_r$ from the training split $\mathbb{S}_t$. To do so, we first initialize the selection set by choosing the image with the maximum number of pixels according to the mask. Next, for each of the remaining image, we compute its rotational geodesic distance to all the selected reference image, and choose the remaining frame based on:

$$i^* = \underset{i\in\mathbb{S}_t, i\notin\mathbb{S}_r}{\operatorname{argmax}} \left(\min_{j\in\mathbb{S}_r} D(\boldsymbol{R}_i, \boldsymbol{R}_j) \right), \tag{15}$$

where $D(\cdot,\cdot)$ denotes the geodesic distance on $\mathbb{SO}(3)$. We repeat the process until enough number of reference images is obtained, which is typically set to 16 following [21].

For applications in the wild when the ground truth object pose is not readily available, we can leverage off-the-shelf SLAM algorithms [51, 56, 66] to compute the poses from the video. Please refer to our supplemental video for relevant results.

5.3. Limitations

Similar to related works [2, 18, 21, 31, 55, 71], our approach focuses on 6D pose estimation and tracking, and relies on external 2D detection, which is obtained from methods such as CNOS [45], or Mask-RCNN [17]. We observe false or missing detection frequently bottlenecks the 6D pose estimation. In future work, an end-to-end framework for novel object detection, 6D pose estimation and tracking would be of interest. Additionally, another typical failure mode due to a combination of multiple challenges is illustrated in Fig. 10.

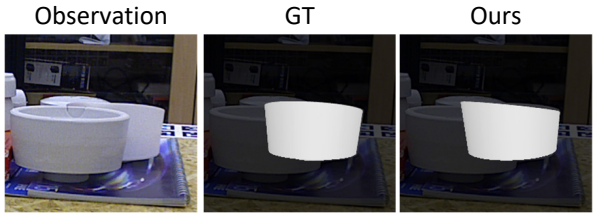

Figure 10. Failure mode. Under the combination of multiple challenges including texture-less, severe occlusion, and limited edge cues, our method fails to estimate the correct orientation.

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 *13th European Conference on Computer Vision (ECCV)*, pages 536–551, 2014. 6, 7

[2] Dingding Cai, Janne Heikkilä, and Esa Rahtu. OVE6D: Object viewpoint encoding for depth-based 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6803–6813, 2022. 5, 7, 3

[3] Ming Cai and Ian Reid. Reconstruct locally, localize globally: A model free method for object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3153–3163, 2020. 2

[4] Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, and Kangxue Yin. TexFusion: Synthesizing 3D textures with text-guided image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4169–4181, 2023. 2, 3

[5] 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 International Conference on Computer Vision (CVPR)*, pages 11973–11982, 2020. 1, 2

[6] 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 (CVPR)*, pages 13142–13153, 2023. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 3

[8] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. PoseRBPF: A Rao-Blackwellized particle filter for 6D object pose tracking. In *Robotics: Science and Systems (RSS)*, 2019. 1, 2, 7, 8

[9] 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. In *International Conference on Learning Representations (ICLR)*, 2021. 5

[10] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. In *International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022. 2

[11] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, pages 3789–3799, 2020. 4

[12] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019. 2

[13] John C Hart. Sphere tracing: A geometric method for the antialiased ray tracing of implicit surfaces. *The Visual Computer*, 12(10):527–545, 1996. 5

[14] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6749–6758, 2022. 7

[15] Poly Haven. Poly Haven: The public 3D asset library. https://polyhaven.com/, 2023. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5, 2

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2961–2969, 2017. 7, 3

[18] 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 (NeurIPS)*, 35: 35103–35115, 2022. 1, 2, 6, 7, 3

[19] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11632–11641, 2020. 1, 2

[20] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. FFB6D: A full flow bidirectional fusion network for 6D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3003–3013, 2021. 1, 2

[21] Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. FS6D: Few-shot 6D pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6814–6824, 2022. 2, 3, 6, 7, 8

[22] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *International Conference on Computer Vision (ICCV)*, pages 858–865, 2011. 6

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020. 2

[24] 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 *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888, 2017. 6, 7

[25] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal,

Stephan Ihrke, Xenophon Zabulis, et al. BOP: Benchmark for 6D object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018. 2, 6

[26] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Third International Workshop on Similarity-Based Pattern Recognition (SIMBAD)*, pages 84–92, 2015. 6

[27] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: Registration of 3D point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4267–4276, 2021. 6

[28] Jan Issac, Manuel Wüthrich, Cristina Garcia Cifuentes, Jeannette Bohg, Sebastian Trimpe, and Stefan Schaal. Depth-based object tracking using a robust gaussian filter. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 608–615, 2016. 1, 2, 7, 8

[29] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018. 1

[30] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *European Conference on Computer Vision (ECCV)*, pages 574–591, 2020. 1, 2

[31] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D pose estimation of novel objects via render & compare. In *6th Annual Conference on Robot Learning (CoRL)*, 2022. 1, 2, 5, 7, 3

[32] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 1

[33] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. TTA-COPE: Test-time adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21285–21295, 2023. 1, 2

[34] Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. NeRF-Pose: A first-reconstruct-then-regress approach for weakly-supervised 6D object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2123–2133, 2023. 2

[35] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep iterative matching for 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 1, 2, 7, 8

[36] Zhigang Li, Gu Wang, and Xiangyang Ji. CDPN: Coordinates-based disentangled pose network for real-time RGB-based 6-DoF object pose estimation. In *CVF International Conference on Computer Vision (ICCV)*, pages 7677–7686, 2019. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *13th European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 3

[38] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Keypoint-based category-level object pose tracking from an RGB sequence with uncertainty estimation. In *International Conference on Robotics and Automation (ICRA)*, 2022. 1, 2

[39] 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. *ECCV*, 2022. 1, 2, 6, 7

[40] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 4

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

[42] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 4

[43] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 4, 1

[44] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3D object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6771–6780, 2022. 5, 6, 7

[45] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023. 1, 3

[46] Brian Okorn, Qiao Gu, Martial Hebert, and David Held. Zephyr: Zero-shot pose hypothesis rating. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 14141–14148, 2021. 7

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 6

[48] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7668–7677, 2019. 1, 2

[49] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. LatentFusion: End-to-end differentiable reconstruction

and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10710–10719, 2020. 2, 7

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[51] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[52] 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 (CVPR)*, pages 6835–6844, 2022. 1, 2, 7

[53] Manuel Stoiber, Martin Sundermeyer, and Rudolph Triebel. Iterative corresponding geometry: Fusing region and depth for highly efficient 3D tracking of textureless objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6865, 2022. 1, 2, 7, 8

[54] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, 2021. 6

[55] 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 (CVPR)*, pages 6825–6834, 2022. 1, 2, 6, 7, 3

[56] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 16558–16569, 2021. 2

[57] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 530–546, 2020. 1, 2

[58] 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 (CoRL)*, pages 306–316, 2018. 2

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 5, 6

[60] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066, 2020. 1, 2

[61] 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 International Conference on Computer Vision (CVPR)*, pages 2642–2651, 2019. 1, 2

[62] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

[63] Bowen Wen and Kostas Bekris. BundleTrack: 6D pose tracking for novel objects without instance or category-level 3D models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074, 2021. 2, 7

[64] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se(3)-TrackNet: Data-driven 6D pose tracking by calibrating image residuals in synthetic domains. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10367–10373, 2020. 1, 2, 6, 7, 8

[65] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. CatGrasp: Learning category-level task-relevant grasping in clutter from simulation. In *International Conference on Robotics and Automation (ICRA)*, pages 6401–6408, 2022. 1

[66] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan 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 (CVPR)*, pages 606–617, 2023. 2, 4, 7

[67] Manuel Wüthrich, Peter Pastor, Mrinal Kalakrishnan, Jeannette Bohg, and Stefan Schaal. Probabilistic object tracking using a range camera. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3195–3202, 2013. 1, 2, 7, 8

[68] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 2, 6, 7

[69] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2492–2502, 2020. 4

[70] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. SSP-Pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7452–7459, 2022. 1, 2

[71] Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi Yang, and Shiliang Pu. Learning symmetry-aware geometry correspondences for 6D object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14045–14054, 2023. 7, 3

[72] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin

Chang. HS-Pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17163–17173, 2023. 2

[73] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019. 5