

3D-Object Perception Transformer (3PT)

Supplementary Material

6. Supplementary Materials

The supplement provides additional details and videos organized.

- *Video versions of Fig. 6 for additional methods*
 - [3pt_vs_Freeze.mp4](#)
 - [3pt_vs_Foundationpose.mp4](#)
 - [3pt_vs_FRTPose.mp4](#)
- *Robotic Bin-Picking:*
 - Sec. 6.1
 - [bin_picking.mp4](#)
- *AR Tracking:*
 - Sec. 6.8
 - [ar_tracking.mp4](#)
- *DIMM Insertion:*
 - Sec. 6.2
 - [dimm.mp4](#)
- *Additional Runtime Discussion:* Sec. 6.3
- *Additional Limitations Discussion:* Sec. 6.4
- *Full Segmentation Results:* Sec. 6.5
- *Full CAD-Free Detection Results:* Sec. 6.6
- *Additional Dataset Details:* Sec. 6.7

6.1. Bin Picking

Setup. A UR10e 6-axis manipulator, equipped with a vacuum suction gripper, is used for emptying objects from a bin. Three overhead cameras are mounted at a distance of 2m that provide grayscale images. *Main Test:* The task involves emptying a bin containing 100 thin sheet-metal parts, each with dimensions of $140 \times 140 \times 1.5$ mm. The parts are randomly scattered in the bin to create a high-density, cluttered environment. *Stress Test:* We extend the task to contain 3 different object classes. 2 of them are very similar looking brackets, only different in number of holes and size.

Challenges. The successful execution of the bin-picking task depends critically on achieving highly accurate 6D pose estimation under challenging geometric and operational constraints. This is compounded by two main factors: long-range perception from the 2m camera height, and severe inter-part occlusion created by the dense, scattered stack of thin (1.5mm) sheet-metal objects. Even minor pose errors, on the order of a few millimeters or degrees, lead to vacuum pressure loss in the gripper or collisions with surrounding objects during extraction. These failures often result in the object being dropped outside the bin or unpredictably thrown outside the bin at high speeds.

We further stress-test the system’s ability to distinguish between brackets with subtle difference such as small changes in geometry (different holes) during the picking task.

Method. We start with an offline grasp planner, which pre-computes valid grasp poses relative to the object’s mesh origin. We then employ an online sense-plan-act strategy for every pick and place cycle. During the sense phase, 3PT generates pairs of 6D pose and scores for each detection. The plan phase then selects the detected object pose associated with the highest score and computes a collision-free grasp and extraction trajectory. The planned motion is executed in the act phase *without* any corrective force, move until suction, or visual servoing feedback. For placement, the objects are simply dropped at a predetermined location.

Results. In the main test, 3PT achieves a 100% bin clear rate, clearing 100/100 parts in the bin with zero failures. For the stress test, it is able to correctly identify, pick, and sort all 3 part types, including similar looking brackets.

6.2. Application: Precision DIMM Insertion

Setup. We employ a KUKA KR6 6-axis industrial robotic arm equipped with a custom two-jaw parallel gripper to manipulate DIMM modules. The perception system consists of three RGB Basler Ace cameras mounted directly on the gripper, arranged 120 degrees apart. The robotic control loop is managed via proprietary software, while 3PT provides the underlying 6DoF pose estimation for both the DIMM modules and the target PCB.

Challenges. The task imposes strict accuracy requirements due to the tight mechanical tolerances of the DIMM and insertion slot. Grasping the DIMM requires a pose accuracy of approximately ± 2 mm from a distance of 30–40cm. The insertion phase is significantly more demanding: the DIMM slot clearance is merely 0.57mm, requiring a PCB pose estimation accuracy within ± 0.285 mm and $\pm 2^\circ$.

Method. The operation follows a coarse-to-fine visual servoing strategy. First, 3PT estimates the PCB pose from a distance of 70cm (where the entire board is visible) to establish a coarse global alignment. Next,

the robot moves to the feeder tray to resolve the 6DoF pose of the incoming DIMM stick for grasping.

For the insertion phase, the strategy varies by slot location due to visibility constraints. For the central 8 slots, the robot hovers at close range to perform a high-precision refinement step immediately prior to each insertion. For the outer 8 slots, where gripper geometry occludes two cameras during the approach, we perform a single refinement step before picking the DIMMs and execute the insertion "blind" based on this cached pose estimate. Crucially, the insertion trajectory relies exclusively on these visual pose estimates without force guidance; force feedback is employed solely for the final locking mechanism (clicking the DIMM into place).

Results. We evaluated the system by filling a complete server tray (16 slots) over 6 independent trials, achieving 96/96 successful insertions. To test robustness, both the DIMM feeder and the receiving PCB were arbitrarily repositioned between runs. A video demonstrating a full sequence of 16 successful insertions is included in the supplementary material.

6.3. Additional Runtime Discussion

Table 3 presents a detailed runtime breakdown across the BOP benchmark suite. We report the average per-scene runtime for our method (3PT) alongside relevant baselines. We note that direct comparisons are complicated by hardware heterogeneity (e.g., H100 vs. RTX 4090) and differing input modalities (RGB vs. RGB-D).

However, several key trends emerge regarding the computational characteristics of 3PT:

Scaling with No. Parts and Image Size. 3PT's detection runtime is primarily influenced by the image resolution (due to our sliding-window tiling strategy) and the size of the 3D object library. This is evident in *HOT3D*, where high-resolution inputs (2048×2048) and a larger object vocabulary (33 parts) result in higher detection latency (10.9s) compared to compact datasets like *LMO* (0.7s). In our demos we only run for a single-object, so runtimes are much faster.

Efficiency vs. Precision Trade-off. While 3PT is slower than lightweight detectors like MUSE [10], this computational cost enables the 3D-conditioned reasoning required for state-of-the-art zero-shot generalization. Crucially, on the challenging *BOP-Industrial* datasets, 3PT is significantly faster than the top-performing baseline, FRTPose [56]. On *XYZ* and *ITODD-MV*, 3PT reduces runtime by a factor of roughly $4\times$ to $8\times$ compared to FRTPose, while relying solely on MV-RGB inputs rather than RGB-D.

Standardized Evaluation. We emphasize that the reported runtimes for 3PT utilize a fixed set of hyperpa-

rameters (e.g., multiple scale priors, unknown classes, top- K hypotheses) across all datasets within a category (e.g. Industrial) to ensure a fair, "out-of-the-box" evaluation. In practical deployments, these thresholds can be tuned to significantly accelerate inference. For example, in bin-picking, we use a single-object class, a single scale prior, a limited ROI, and only 3 cameras.

Code Optimizations Runtime can further be improved without algorithmic changes. Some of the code is not optimally implemented, for example the refinement could cache the image features across hypotheses since we run the same number of hypotheses across images.

6.4. Additional Limitations

This section outlines limitations related to input modalities and architectural design.

- RGB-D Integration:** While 3PT is designed to eliminate the dependency on depth sensors, there are scenarios where RGB-D is the only available modality or where depth data is high-quality. Currently, our framework does not leverage depth information. Extending the architecture to fuse depth features during the refinement stage could further enhance performance when multi-view is unavailable, and remains a promising direction for future work.
- Unified Architecture:** Although 3PT streamlines the pipeline compared to ensemble-based methods, it currently utilizes two separate models for detection (3PT-D) and refinement (3PT-R). A fully end-to-end trainable framework that unifies these stages into a single model could theoretically reduce feature redundancy and optimize inference efficiency.
- Texture Gap in 3D Models:** 3PT-D and 3PT-R are conditioned on renders of 3D object models. Consequently, performance degrades when there is a visual discrepancy between the provided CAD model and the physical object (e.g., texture mismatches or geometric simplifications), as observed in subsets of the IC-BIN dataset. While we achieve state-of-the-art performance on average, addressing these specific edge cases via render data augmentations during training is an area for future exploration. This limitation of being overly sensitive to subtle differences, however, can also be encouraged for certain applications. For instance, when sorting multiple similar objects, our approach is less prone to confusing them. A capability we demonstrate in [bin_picking.mp4](#).

6.5. Full Segmentation Results

As described in Sec. 3.3, 3PT-R generates per-instance 2D segmentation masks alongside the refined poses. On the BOP-Classic benchmark suite (Tab. 4), 3PT achieves superior average performance compared to

Table 3. Dataset statistics and runtime comparison. Runtimes are reported in seconds.

Metric	BOP-Classic							BOP-H3			BOP-Industrial		
	ITODD	LMO	YCB-V	T-LESS	HB	IC-BIN	TUD-L	HOT3D	HANDAL	HOPEv2	IPD	XYZ	ITODD-MV
No. Parts	28	8	21	30	33	2	3	33	40	28	10	15	28
No. Scenes	721	200	900	1000	300	150	600	5140	1684	457	1232	60	721
Image Size	960x1280	480x640	480x640	540x720	480x640	480x640	480x640	2048x2048	1440x1920	1080x1920	1544x2064	1080x1440	960x1280
RGB	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
RGB-D	✓	✓	✓	✓	✓	✓	✓	×	×	✓	✓	✓	✓
MV-RGB	×	×	×	×	×	×	×	✓	×	×	✓	✓	✓
<i>Detection Runtime (seconds)</i>													
3PT (H100)	2.631	0.737	1.876	2.704	2.989	0.196	0.278	10.987	4.131	4.536	2.092	2.040	2.631
MUSE (4090)	0.406	0.585	0.446	0.519	0.465	0.556	0.559	0.722	1.038	0.995	0.688	0.671	0.406
<i>6DoF Pose Runtime (seconds)</i>													
3PT (H100)	6.941	4.768	5.261	14.079	10.043	2.443	1.540	18.224	14.522	15.292	24.516	42.138	24.979
Co-Op (4090)	3.368	12.113	6.313	9.295	9.090	11.352	9.561	1.958	3.237	4.864	—	—	—
FRTPose (5090)	—	—	—	—	—	—	—	—	—	—	33.897	160.484	230.962
FreeZEV2.2 (L40)	—	—	—	—	—	—	—	—	—	—	7.201	22.811	20.475

Table 4. 2D segmentation results for the BOP-Classic evaluation suite. While 3PT is the best algorithm on average, the majority of the improvement stems from the T-Less subset. We report Average Precision in %.

Method	LMO[4]	T-LESS[23]	TUD-L[26]	IC-BIN[15]	ITODD[17]	HB[31]	YCB-V[50]	Avg
CNOS [33]	39.7	37.4	48.0	27.0	25.4	51.1	59.9	41.2
SAM-6D [35]	46.0	45.1	56.9	35.7	33.2	59.3	60.5	48.1
MUSE [10]	47.7	47.8	57.3	43.3	39.1	63.5	69.0	52.5
Ours	47.4	64.0	56.0	31.9	43.1	68.0	64.7	53.6
Δ	-0.3	+16.2	-1.3	-11.4	+4.0	+4.5	-4.3	+1.1

Table 5. Comparison of detection methods. We report the Average Precision in %. Δ denotes the raw percentage point difference w.r.t. the best competitor.

Method	HOT-3D	HOPEv2	HANDAL	Average
CNOS (FastSAM)	37.3	34.3	30.4	34.0
CNOS 25 (YOLOE-DINOv3)	48.1	45.2	38.9	44.1
Ours	45.8	56.8	41.9	48.2
Δ	-2.3	+11.6	+3.0	+4.1

prior methods, though this improvement is largely attributed to its performance on the T-Less [23] subset.

6.6. Full Model-Free Detection Results

Model-Free Detection describes the scenario for 2D object detection (bounding box) where the object’s CAD model is not provided as an input to the system. This setting is common when CAD models are proprietary or unavailable. The system is instead provided with a gallery of input images (photographs) of the object, typically captured from viewpoints relevant to the application domain. We assume binary segmentation masks are available for each of these object images (as is available in the BOP challenge).

The 3PT-D framework is readily extendable to this model-free setting. As 3PT’s interaction with the CAD model is exclusively through the rendering pipeline, we adapt the framework by replacing any required synthetic rendering operation. Specifically, at each step where a render is needed, 3PT-D is modified to randomly select one of the provided masked photographs as the input.

The performance of this model-free adaptation is summarized in Tab. 5 on BOP-H3. BOP-Classic and BOP-Industrial do not have model-free benchmarks. We demonstrate significant performance gains on two out of the three evaluation datasets. Specifically, we achieve an 11.6 percentage point improvement on HOPEv2 and exhibit 4.1 increase on metric averaged across all tested datasets.

6.7. Dataset Curation

As mentioned in the paper, we trained the networks on 900,000 unique synthetic images, using over 100,000 mesh files. The dataset has approximately 100 million unique trainable object instances before data augmentation.

The CAD models were sourced from many different public datasets, including Objaverse[13], ABO[12], GSO[16], and OmniObject3D[59]. In addition to these meshes, we also use a different set of CAD models as distractor objects to help the network avoid false positive detections and to increase the diversity of the dataset.

For each image, we arranged the objects in one of three different ways: randomly oriented in mid-air, dropped onto a flat surface, and tightly packed together. The first arrangement ensures that all sides of the network would see the objects from a wide variety of angles. The second arrangement ensures that the data includes realistic occlusions and physically realistic poses. And the third arrangement helps the network differentiate between two objects extremely close together. It also produces realistic occlusions and simulates certain common real world scenarios.

Another important element being randomized are the material properties of the CAD models. For each instance, we apply a small jitter to its metallicness, roughness, and color properties. This helps the model be more robust to differences between the CAD prompts



Figure 7. An example of a scene where objects have been dropped onto a surface into a dense pile.

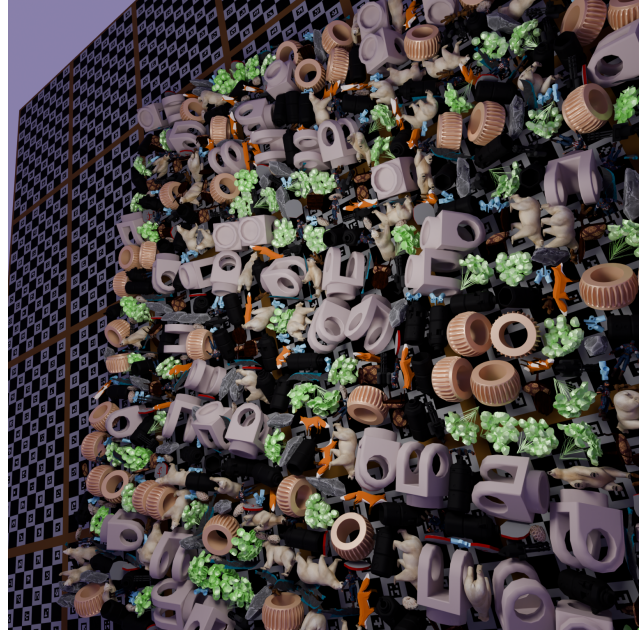


Figure 8. An example of a scene where objects are tightly packed together.

and the images of the object in the real world.

The dataset also features a wide variety of lighting conditions. In many of the scenes, we place directional and/or non-directional light sources with various colors and intensities. This simulates difficult light-related challenges found in the real world, such as shadows, non-uniform lighting, and extremely bright/dark conditions. To improve the realism of the lighting, we also leverage HDRI environments. Finally, we use Blender's ray tracer to further improve the photorealism of the images.

6.8. Applications: AR Tracking

This section details the implementation and demo for the Augmented Reality (AR) tracking application introduced in Sec. 4.3 and Fig. 4(a) of the main paper.

We visualize our method on the *HOT3D* [1] dataset, which captures egocentric interaction scenarios using AR glasses. This dataset presents unique perception challenges: it relies on multi-view RGB (2 or 3 cameras depending on the device) with severe fisheye distortion and completely lacks depth sensor data [1]. The data consists of 5-second clips recorded at 30 fps (150 frames total).

To address this, we implement a detect-and-track pipeline. We employ the full 3PT framework to detect and estimate the initial object pose on the first frame ($t = 0$). For the subsequent 149 frames, we bypass the detection stage and exclusively utilize the refinement module (3PT-R) to update the pose. Specifically,

we run a single iteration of 3PT-R on the multi-view inputs, initialized with the pose from the previous time step. This lightweight update operates at approximately 35ms per frame on an NVIDIA H100 GPU. The resulting high-stability tracking is visualized in the supplementary video *AR_Tracking.mp4*.

References

- [1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. *CVPR*, 2025. 6, 7, 8, 4
- [2] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 3
- [3] Matteo Bortolon, Theodore Tsesmelis, Stuart James, Fabio Poiesi, and Alessio Del Bue. 6dgs: 6d pose estimation from a single image and a 3d gaussian splatting model. In *European Conference on Computer Vision (ECCV)*, 2024. 3
- [4] Eric Brachmann. 6D Object Pose Estimation using 3D Object Coordinates [Data], 2020. 7, 3
- [5] Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *ECCV*, pages 414–431. Springer, 2024. 2, 3, 7, 8
- [6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4
- [7] Long Chen, Han Yang, Chenrui Wu, and Shiqing Wu. Mp6d: An rgb-d dataset for metal parts 6d pose estimation. *IEEE Robotics and Automation Letters*, 7(3):5912–5919, 2022. 7
- [8] Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pages 381–396. Springer, 2022. 7
- [9] Sungmin Cho, Sungbum Park, and Insoo Oh. F3dt2d: Method submission to the bop challenge 2024. https://bop.felk.cvut.cz/method_info/761/, 2024. BOP: Benchmark for 6D Object Pose Estimation. 2
- [10] Sungmin Cho, Sungbum Park, and Insoo Oh. Muse: Model-based uncertainty-aware similarity estimation for zero-shot 2d object detection and segmentation. *arXiv preprint arXiv:2510.17866*, 2025. 2, 7, 8, 3
- [11] Alvaro Collet and Siddhartha S Srinivasa. Efficient multi-view object recognition and full pose estimation. In *2010 IEEE International Conference on Robotics and Automation*, pages 2050–2055. IEEE, 2010. 5
- [12] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21126–21136, 2022. 6, 3
- [13] 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. 6, 3
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 4
- [15] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7, 3
- [16] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560, 2022. 6, 3
- [17] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017. 2, 6, 7, 8, 3
- [18] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130): 295–305, 1953. 5
- [19] Bernd Von Gimborn, Philipp Ausserlechner, Markus Vincze, and Stefan Thalhammer. Diffusion features for zero-shot 6dof object pose estimation, 2024. 3
- [20] Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *IROS*, 2023. 6, 7, 8
- [21] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022. 3
- [22] 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 *Computer Vision – ACCV 2012*, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 6
- [23] Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, 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, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 7, 3
- [24] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision (ECCV)*, 2018. 6
- [25] Tomáš Hodaň et al. BOP: Benchmark for 6d object pose estimation leaderboards. <https://bop.felk.cvut.cz/leaderboards/>, 2025. Accessed: 2025-11-13. 7
- [26] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *Computer Vision ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, page 1935, Berlin, Heidelberg, 2018. Springer-Verlag. 7, 3
- [27] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 3, 7
- [28] Junwen Huang, Jizhong Liang, Jiaqi Hu, Martin Sundermeyer, Peter KT Yu, Nassir Navab, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity, 2025. 2, 6, 7, 8
- [29] Junwen Huang, Shishir Reddy Vutukur, Peter KT Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Raypose: Ray bundling diffusion for template views in unseen 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9102–9112, 2025. 3, 7
- [30] Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, and Michael Stark. Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22691–22701, 2024. 6, 7, 8
- [31] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects . In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2767–2776, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 7, 3
- [32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, and et al. Segment anything. In *ICCV*, 2023. 1
- [33] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 3
- [34] 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 *Conference on Robot Learning (CoRL)*, 2022. 3
- [35] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui 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. 2, 3, 7
- [36] Lihua Liu, Jiehong Lin, Zhenxin Liu, and Kui Jia. Picopose: Progressive pixel-to-pixel correspondence learning for novel object pose estimation. *arXiv preprint arXiv:2504.02617*, 2025. 3, 7
- [37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2
- [38] Yangxiao Lu, Jishnu Jaykumar P, Yunhui Guo, Nicholas Ruozzi, and Yu Xiang. Adapting pre-trained vision models for novel instance detection and segmentation. *arXiv preprint arXiv:2405.17859*, 2024. 2
- [39] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, pages 728–755. Springer, 2022. 4
- [40] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. 36:72983–73007, 2023. 4
- [41] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 7
- [42] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Co-op: Correspondence-based novel object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 6, 7, 8
- [43] Van Nguyen Nguyen, Thibault Groueix, Georgy Pomiatkin, 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. 2, 7, 3
- [44] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent 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*, 2024. 3, 7
- [45] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sung-

- 1062 phill Moon, Hyeontae Son, Lukas Ranftl, Jonathan
1063 Tremblay, Eric Brachmann, Bertram Drost, Vincent
1064 Lepetit, Carsten Rother, Stan Birchfield, Jiri Matas,
1065 Yann Labbe, Martin Sundermeyer, and Tomas Hodan.
1066 BOP challenge 2024 on model-based and model-free 6D
1067 object pose estimation. *IEEE Conference on Computer
1068 Vision and Pattern Recognition Workshops (CVPRW,
1069 CV4MR Workshop)*, 2025. 3
- [46] Maxime Oquab, Timothée Darcet, Théo Moutakanni,
1070 Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre
1071 Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin
1072 El-Nouby, et al. Dinov2: Learning robust visual
1073 features without supervision. *Transactions on Machine
1074 Learning Research*, 2024. 3, 6
- [47] Lukas Ranftl, Felix Brendel, Bertram Drost, and
1075 Carsten Steger. Mvtop: Multi-view transformer-
1076 based object pose-estimation. *arXiv preprint
1077 arXiv:2508.03243*, 2025. 3
- [48] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun.
1078 Vision transformers for dense prediction. In *Proceed-
1079 ings of the IEEE/CVF International Conference on
1080 Computer Vision (ICCV)*, pages 12179–12188, 2021. 6
- [49] Hamid Rezaatoughi, Nathan Tsoi, JunYoung Gwak,
1081 Amir Sadeghian, Ian Reid, and Silvio Savarese.
1082 Generalized intersection over union: A metric and a
1083 loss for bounding box regression. In *Proceedings of
1084 the IEEE/CVF Conference on Computer Vision and
1085 Pattern Recognition (CVPR)*, pages 658–666, 2019. 4
- [50] Tim Salzmann, Markus Ryhl, Alex Bewley, and
1086 Matthias Minderer. Scene-graph vit: End-to-end
1087 open-vocabulary visual relationship detection. In
1088 *ECCV*, pages 195–213. Springer, 2024. 4
- [51] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan
1089 Ilic. Osop: A multi-stage one shot object pose
1090 estimation framework, 2022. 3
- [52] Ritvik Singh, Jingzhou Liu, Karl Van Wyk, Yu-Wei
1091 Chao, Jean-Francois Lafleche, Florian Shkurti, Nathan
1092 Ratliff, and Ankur Handa. Synthetica: Large scale
1093 synthetic data for robot perception, 2024. 6
- [53] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He,
1094 Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou.
1095 Onepose: One-shot object pose estimation without cad
1096 models. In *Proceedings of the IEEE/CVF Conference
1097 on Computer Vision and Pattern Recognition*, pages
1098 6825–6834, 2022. 3
- [54] Stephen Tyree, Jonathan Tremblay, Thang To, Jia
1099 Cheng, Terry Mosier, Jeffrey Smith, and Stan Birch-
1100 field. 6-dof pose estimation of household objects
1101 for robotic manipulation: An accessible dataset and
1102 benchmark. In *International Conference on Intelligent
1103 Robots and Systems (IROS)*, 2022. 6, 7, 8
- [55] Jianyuan Wang, Christian Rupprecht, and David
1104 Novotny. Posediffusion: Solving pose estimation via
1105 diffusion-aided bundle adjustment. In *Proceedings of
1106 the IEEE/CVF International Conference on Computer
1107 Vision (ICCV)*, 2023. 3
- [56] Yulin Wang, Mengting Hu, Jianghao Zhuo,
1108 and Chen Luo. Frt-pose wapr2 (multi-cam).
1109 https://bop.felk.cvut.cz/method_info/1157/,
1110 2025. Accessed: 2025-10-27. 2, 3, 7, 8
- [57] Bowen Wen, Chaitanya Mitash, and Kostas Bekris.
1111 Data-driven 6d pose tracking by calibrating im-
1112 age residuals in synthetic domains. *arXiv preprint
1113 arXiv:2105.14391*, 2021. 7
- [58] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield.
1114 Foundationpose: Unified 6d pose estimation and track-
1115 ing of novel objects. In *CVPR*, pages 17868–17879,
1116 2024. 3, 7
- [59] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei
1117 Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang,
1118 Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d:
1119 Large-vocabulary 3d object dataset for realistic percep-
1120 tion, reconstruction and generation. In *Proceedings of
1121 the IEEE/CVF Conference on Computer Vision and
1122 Pattern Recognition (CVPR)*, pages 803–814, 2023. 6, 3
- [60] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan,
1123 and Dieter Fox. Posecnn: A convolutional neural
1124 network for 6d object pose estimation in cluttered
1125 scenes. 2018. 6, 7, 3
- [61] Jun Yang, Yizhou Gao, Dong Li, and Steven L Waslan-
1126 der. Robi: A multi-view dataset for reflective objects in
1127 robotic bin-picking. In *2021 IEEE/RSJ International
1128 Conference on Intelligent Robots and Systems (IROS)*,
1129 pages 9788–9795. IEEE, 2021. 2
- [62] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du,
1130 Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast
1131 segment anything. *arXiv preprint arXiv:2306.12156*,
1132 2023. 2
- [63] Evin Pnar Örneek, Yann Labbé, Bugra Tekin, Lingni
1133 Ma, Cem Keskin, Christian Forster, and Tomas Ho-
1134 dan. Foundpose: Unseen object pose estimation with
1135 foundation features, 2024. 3, 7