Pose Estimation of Sub-Centimeter Industrial Parts for Automated Assembly

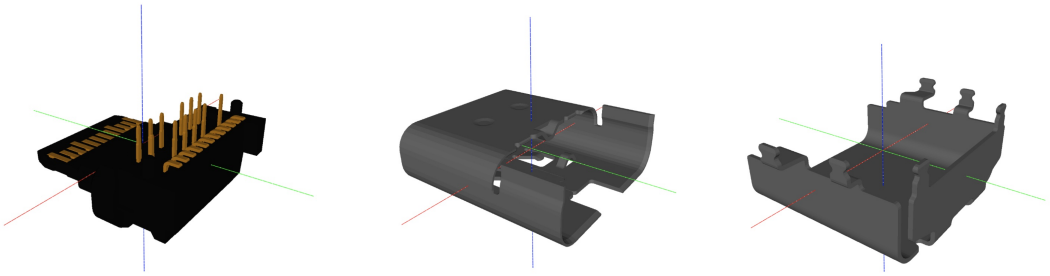
Hameed Abdul-Rashid[1], **Yangfei Dai**[1], **Holly Dinkel**[1], **Minh Quang Ta**[2], **Tan Chen**[3], **Jungyi Geng**[4], and **Timothy Bretl**[1]

[1]**University of Illinois Urbana-Champaign**
[2]**University of West Florida**
[3]**Michigan Technological University**
[4]**Pennsylvania State University**
[1]{hameeda2,yangfei4,hdinkel2,tbretl}@illinois.edu, [2]mta@uwf.edu,
[3]tanchen@mtu.edu, [4]jgeng@psu.edu

Abstract: This work presents a learned, vision-based system for estimating the pose of three sub-centimeter parts for automating a high-mix low-volume assembly line. In this system, a 3D model of each part is input to a BlenderProc2 physics-based rendering engine to produce a photorealistic synthetic dataset of part images. The synthetic images are used to train a Mask R-CNN model to segment part instances in a scene, with instance mask labels used during training obtained automatically without manual labeling. Instance segmentation of individual parts enables selection for assembly when multiple parts are present. Finally, a PVNet model is trained on individual parts instance image crops to estimate planar part pose within 0.6 mm, 0.59 mm, 63.81 mm XYZ positional error and 4.08° angular error. An ablation study injecting noise into the keypoint detection stage of the PVNet-based model is performed to evaluate robustness of pose estimation to sensor noise. This system offers a cost-effective approach to learning sub-centimeter pose estimation.

Keywords: Pose Estimation, Object Segmentation, Industrial Assembly, Robotic Manipulation

1 Introduction



(a) Inserted Part (b) Main Part (c) Top Part

Figure 1: The CAD models of three sub-centimeter, symmetric industrial parts, are inputs to the learned pose estimation system.

The objective of this work is to estimate the 6-DOF pose of three sub-centimeter parts–a main part, a top part and an inserted part –which are assembled to form a USB Type-C connector. Since automation by custom machine design is not cost-effective in low volume production, human operators currently assemble these parts manually. Accurate pose estimation of these sub-centimeter parts is an enabling step in high-precision sub-mm tolerance assembly by autonomous industrial robots in high-mix, low-volume production lines [1]. Other work investigates small-scale part assembly with

8th Conference on Robot Learning (CoRL 2024), Munich, Germany.

robot arms in simulation but assumes known object poses prior to grasp [2]. This paper presents a vision-based learning system that enables pose estimation of industrial parts for assembly.

The accessibility of computer vision models and robotic hardware is nowadays unprecedented. With this accessibility, comes the unspoken assumption that off-the-shelf models and hardware can easily be fine-tuned and integrated for various applications. When the performance of pose estimation and grasp success rate impacts the bottom line of a highly controlled and automated industrial settings, how well contemporary off-the-shelf models work in practice comes into question.

Contemporary datasets focus on everyday objects that have distinctive visual features [3, 4, 5]. Figure 1 shows the the symmetric and reflective surfaces of three sub-centimeter USB Type C components. While there are datasets that contain objects with similar symmetric and reflective properties [6, 7, 8, 9, 10], to this paper focuses on sub-centimeter USB Type C components in a high mix low-volume setting for robotic assembly applications which there is small tolerance for error. Specifically, this paper's contributions are:

1. This work explores the sensitivity to pixel noise and object distance of the proposed pipeline's pose estimation result.
2. A demo of this work's automated robotic assembly pipeline that uses pose estimation of sub-centimeter industrial parts in a high mix, low volume scenario. Code is shared at github.com/yangfei4/clean-pvnet.

2 Related Work

2.1 Pose Estimation

Estimating the position and orientation of an object is a prerequisite for robotic assembly [4, 3, 11, 12, 13, 14, 15, 9, 16]. This paper uses a supervised deep learning method to estimate the poses of USB Type-C components. PVNet predicts the segmentation mask of an object and use pixel-wise voting of the network output to determine the location of 2D keypoints [17]. Peng et al. also proposes a custom PnP solver that takes a subset of the predicted 2D keypoints and known set of 3D keypoints from training to estimate the 6D pose of the object. This work adopts PVNet and makes several modifications as detailed in Section 3.

2.2 Object Detection and Instance Segmentation

In the case of having several *instances*, multiple objects of the same type, many contemporary learning-based methods of pose estimation may require instance-specific information [18, 19, 20]. Object detection coarsely denotes which region of the image an instance is in while instance segmentation is more fine grained and classifies specific pixels as belonging to the instance of interest. Depending on the pose estimation method, either object detection or instance segmentation information may be needed as prior information before estimating the full 6DoF pose. For this work, the pose estimation method may need isolated information for each USB Type-C instance. Mask R-CNN is used as the contemporary method of object detection and instance segmentation [21].

3 Method

Figure 2 shows the pose estimation pipeline used in this work. The Mask R-CNN instance segmentation network and the PVNet pose predictor are used to detect and estimate the pose of each part in an image. The Mask R-CNN model predicts the class, mask, and bounding box for each part instance. The center pixel of the part instance bounding box is used to crop an image patch input to PVNet. The PVNet model provides an initial pose estimate based on the cropped image patch. The orientation estimates of PVNet are refined by assuming the part instance lies on a table in a stable planar orientation. The list of stable pose candidates are shown at Figure 3. The z-component of

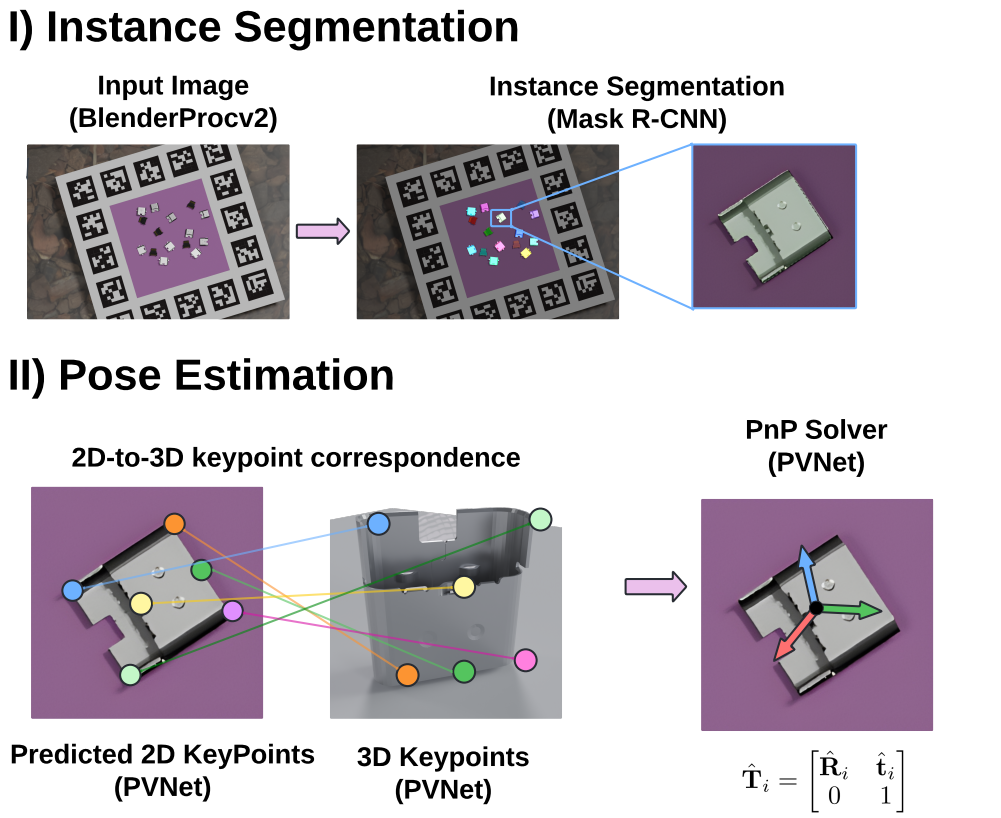


Figure 2: Overview of the proposed pose estimation pipeline.

the position for each part instance is also replaced with the camera-to-(x, y) position on the table obtained from extrinsic calibration. A description of the pipeline is provided in the Appendix.

4 Experiments

4.1 Synthetic Dataset Generation

Figure 4 illustrates the pipeline used to generate training and testing data. Each training dataset contains 20,000 images with a resolution of 256x256 pixels. The annotations provide essential details, including the 6D pose of each object relative to the camera, segmentation masks for each

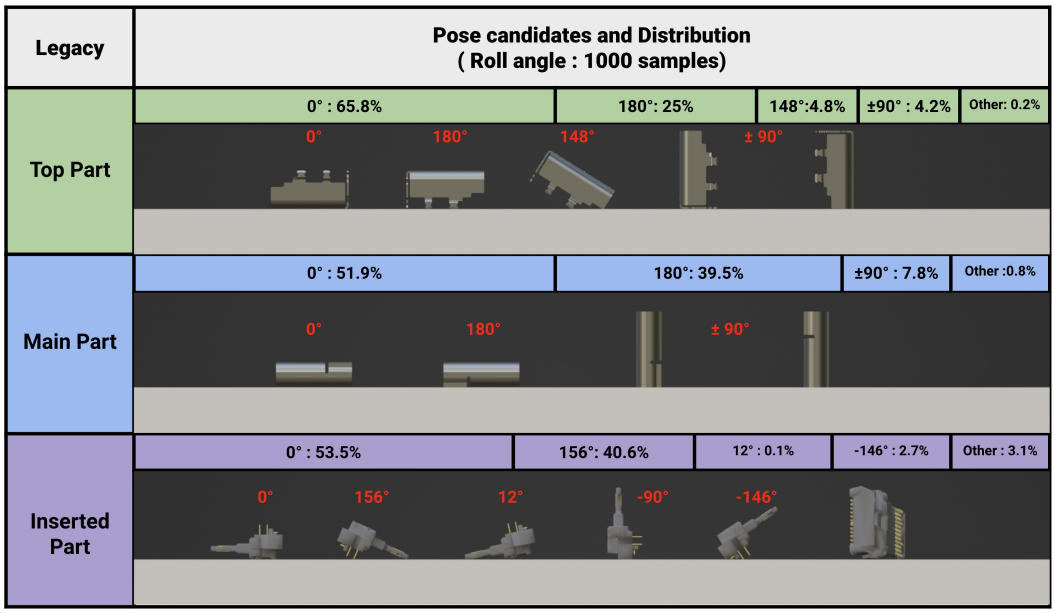


Figure 3: This figure presents the stable pose candidates for each individual part. Traditionally, stable poses are derived from object geometry, which can be inefficient and unreliable. Our pipeline automates this process using the BlenderProc2 simulation engine, capturing all potential stable poses and their associated probabilities. Each object is initialized in random poses within a simulated environment and allowed to fall onto a surface. Once stationary, the final poses are recorded as rotation matrices, repeated until a sufficient number of samples (e.g., 5000 per object) is collected. The rotation matrices are then converted into Euler angles (roll, pitch, and yaw), which describe the object's orientation in 3D space. Roll represents rotation around the object's x-axis, pitch around the y-axis, and yaw around the z-axis. Due to symmetry along the y-axis, only the roll angle is considered, with yaw varying between 0° and 360°.

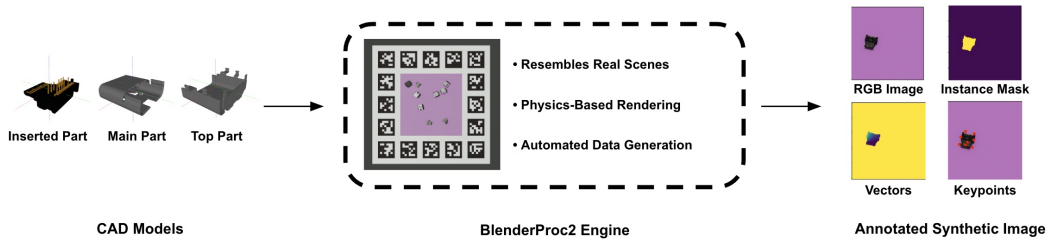


Figure 4: BlenderProc2 is used to render synthetic images from CAD models, replicating real-world scenes. A physics-based simulation ensures objects are placed on a tagboard surface in physically realistic poses. This process automatically generates accurate labels, including instance masks, unit vectors representing the direction from each object pixel to predefined keypoints, and the coordinates of 8 pre-defined keypoints.

object, and the exact locations of eight keypoints. These keypoints are selected through the farthest point sampling (FPS) algorithm.

4.2 Evaluation On Synthetic Dataset

Table 1 presents the raw output of PVNet evaluated on a synthetic testing dataset consisting of 2,000 images. Two metrics are used to assess the accuracy of 6D pose estimation in the context of robotic assembly, i.e. translation and angular (rotation) errors. The predicted pose, $\hat{\mathbf{T}}$, is a transformation

Table 1: PVNet Evaluation Error

Part	**Translation Error** (mm)						**Angular Error** (°)	
	x		y		z		θ	
	μ	$(\pm\sigma)$	μ	$(\pm\sigma)$	μ	$(\pm\sigma)$	μ	$(\pm\sigma)$
Inserted Part	0.60	0.45	0.59	0.42	63.80	17.24	2.74	1.60
Main Part	0.46	0.37	0.47	0.38	51.74	19.38	6.19	2.52
Top Part	0.74	0.54	0.71	0.54	75.90	23.13	3.31	2.09
Average	0.60	0.45	0.59	0.45	63.81	19.92	4.08	2.07

matrix in the world frame,

$$\hat{\mathbf{T}}_i = \begin{bmatrix} \hat{\mathbf{R}}_i & \hat{\mathbf{t}}_i \\ 0 & 1 \end{bmatrix} \tag{1}$$

where $\hat{\mathbf{t}} \in \mathbb{R}^{3\times 1}$ is the predicted i^{th} translation vector and $\hat{\mathbf{R}} \in \mathbb{R}^{3\times 3}$ is the predicted i^{th} rotation matrix. Similarly, the i^{th} true pose is

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ 0 & 1 \end{bmatrix}. \tag{2}$$

Since the object pose is expressed as an orientation and a position, two error metrics are used to quantitatively evaluate the estimate. Assembly of sub-centimeter components with tight insertion tolerances requires accurate pose estimates for picking and assembly. The translation error quantifies the difference between the estimated and true positions in Cartesian space. The translation error is calculated as the Euclidean distance between the predicted translation, $\hat{\mathbf{t}}_i = (\hat{x}_i, \hat{y}_i, \hat{z}_i)$, and the true translation, $\mathbf{t}_i = (x_i, y_i, z_i)$. The translation errors for each dimension are

$$\begin{aligned} e_{x,i} &= \mid \hat{x}_i - x_i \mid \\ e_{y,i} &= \mid \hat{y}_i - y_i \mid . \\ e_{z,i} &= \mid \hat{z}_i - z_i \mid \end{aligned} \tag{3}$$

Angular error measures the deviation in orientation between the predicted and actual poses. The angular error quantifies the difference between the predicted rotation $\hat{\mathbf{R}}$ and the true rotation $\mathbf{R}$ as

$$e_{\theta,i} = \cos^{-1}\left(\frac{\mathrm{tr}(\mathbf{R}_i^{\intercal}\hat{\mathbf{R}}_i) - 1}{2}\right) \tag{4}$$

4.3 Ablation Study - Sensitivity of PnP Solver

This ablation study investigates the sensitivity of the Perspective-n-Point (PnP) solver to errors in keypoint detection. In two-stage methods, such as PvNet, a set of predefined 2D keypoints are estimated from RGB images using supervised learning techniques. The object pose is subsequently determined by establishing 2D-3D correspondences between the input image and a CAD model, minimizing keypoint reprojection errors—typically solved using the PnP algorithm. Our analysis reveals that this optimization process is highly sensitive to inaccuracies in keypoint detection. Specifically, even small pixel shifts or errors in keypoint coordinates can lead to significant pose estimation errors. The degree of this sensitivity is influenced by several factors, including the focal length of the sensor, the distance between the object and the sensor, and the size of the object. We conducted a series of experiments to examine the relationship between pose prediction errors and keypoint noise, with a focus on object-to-image distance, denoted as μ_d. This distance is defined as the average distance between the center of the object and the camera's center, computed over 500 samples for each dataset:

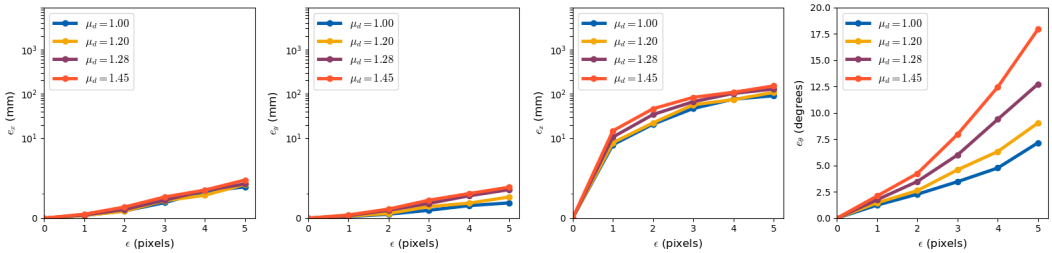


Figure 5: These plots illustrate the relationship between translation prediction errors (first three plots) and rotation prediction error (last plot) as a function of noise level, for different object-to-camera distances (μ_d). Different levels of noise were introduced to the x or y coordinates of eight pre-defined keypoints in the dataset. These perturbed keypoints, along with the camera intrinsics, were then input into an efficient Perspective-n-Point (EPnP) solver [22] to estimate the 6D object pose relative to the camera. The left and center-left plots show that the average translation errors in the x and y directions (e_x and e_y) are relatively insensitive to keypoint localization errors (ϵ). The center-right plot demonstrates that the average z-axis translation error (e_z) exhibits sub-exponential growth as the noise level increases, with larger object-to-camera distances exacerbating the error. The right plot reveals that the average rotation prediction error (e_θ) grows exponentially with increasing noise, and this error is further amplified at greater object-to-camera distances.

$$\mu_d = \frac{1}{n} \sum_{i=1}^{n} d_i \tag{5}$$

where n is the number of samples in the dataset (in this case, 500), d_i represents the distance between the object center and the camera center for the i-th sample.

5 Conclusion

In this work, a vision-based learning system is presented for pose estimation of sub-centimeter industrial parts, enabling automated assembly in a high-mix, low-volume scenario. A synthetic dataset of part images is generated and used to train learning models for part pose estimation. Future work will focus on integrating the developed pose estimation pipeline into a complete assembly line of components.

Acknowledgments

The authors thank the members of the UIUC-FIT CoBot Factory Project and the teams developing the open-source software used in this project. All authors were supported by the Foxconn Interconnect Technology (FIT) and the Center for Networked Intelligent Components and Environments (C-NICE) at the University of Illinois Urbana-Champaign. Holly Dinkel was also supported by the Graduate Assistance in Areas of National Need award P200A180050-19 and the NASA Space Technology Graduate Research Opportunity awards 80NSSC21K1292.

References

[1] T. Chen, Z. Huang, J. Motes, J. Geng, Q. M. Ta, H. Dinkel, H. Abdul-Rashid, J. Myers, Y.-J. Mun, W.-C. Lin, Y.-Y. Yang, S. Liu, M. Morales, N. M. Amato, K. Driggs-Campbell, and T. Bretl. Insights from an Industrial Collaborative Assembly Project: Lessons in Research and Collaboration. In *IEEE Int. Conf. Robot. Autom. (ICRA) Workshop on Collaborative Robots and Work of the Future*, 2022.

[2] M. Q. Ta, H. Dinkel, H. Abdul-Rashid, Y. Dai, J. Myers, T. Chen, J. Geng, and T. Bretl. The Impact of Time Step Frequency on the Realism of Robotic Manipulation Simulation for

Objects of Different Scales. In *IEEE Int. Conf. Intell. Robot. Sys. (IROS) Workshop on Robotics and AI in Future Factory*, 2023.

[3] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and c. Rother. Uncertainty-Driven 6D Pose Estimation of Objects and Scenes From a Single RGB Image. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, June 2016.

[4] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robot. Sci. Syst. (RSS)*, Pittsburgh, Pennsylvania, June 2018.

[5] T. Hodan, M. Sundermeyer, Y. Labbe, V. N. Nguyen, G. Wang, E. Brachmann, B. Drost, V. Lepetit, C. Rother, and J. Matas. BOP Challenge 2023 on Detection Segmentation and Pose Estimation of Seen and Unseen Rigid Objects. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5610–5619, 2024.

[6] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects. In *IEEE Winter Conf. Comput. Vis. (WACV)*, pages 880–888. IEEE, 2017.

[7] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 858–865. IEEE, 2011.

[8] B. Drost, M. Ulrich, P. Bergmann, P. Härtinger, and C. Steger. Introducing MVTec ITODD - A Dataset for 3D Object Recognition in Industry. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pages 2200–2208, 2017.

[9] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield. 6-DoF Pose Estimation of Household Objects for Robotic Manipulation: An Accessible Dataset and Benchmark. In *IEEE/RSJ Int. Conf. Intell. Robot. Sys. (IROS)*, 2022.

[10] J. Yang, Y. Gao, D. Li, and S. L. Waslander. ROBI: A Multi-View Dataset for Reflective Objects in Robotic Bin-Picking. In *IEEE/RSJ Int. Conf. Intell. Robot. Sys. (IROS)*, 2021.

[11] R. Brégier, F. Devernay, L. Leyrit, and J. L. Crowley. Symmetry Aware Evaluation of 3D Object Detection and Pose Estimation in Scenes of Many Parts in Bulk. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017.

[12] J. Yang, D. Li, and S. L. Waslander. Probabilistic multi-view fusion of active stereo depth maps for robotic bin-picking. *IEEE Robotics and Automation Letters*, 6(3):4472–4479, 2021.

[13] F. von Drigalski, K. Hayashi, Y. Huang, R. Yonetani, M. Hamaya, K. Tanaka, and Y. Ijiri. Precise Multi-Modal In-Hand Pose Estimation using Low-Precision Sensors for Robotic Assembly. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021.

[14] Y. Wu, M. Zand, A. Etemad, and M. Greenspan. Vote from the Center: 6 DoF Pose Estimation in RGB-D Images by Radial Keypoint Voting. In *Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[15] P. Wang, H. Jung, Y. Li, S. Shen, R. P. Srikanth, L. Garattoni, S. Meier, N. Navab, and B. Busam. PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation With Photometrically Challenging Objects. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[16] B. Wen, C. Mitash, B. Ren, and K. E. Bekris. SE(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In *IEEE/RSJ Int. Conf. Intell. Robot. Sys. (IROS)*, pages 10367–10373. IEEE, 2020.

[17] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. PVNet: Pixel-wise Voting Network for 6DoF Pose Estimation. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[18] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit. Templates for 3D Object Pose Estimation Revisited: Generalization to New objects and Robustness to Occlusions. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022.

[19] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit. GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[20] G. Wang, F. Manhardt, F. Tombari, and X. Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In *IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 16611–16621, 2021.

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[22] V. Lepetit, F. Moreno-Noguer, and P. Fua. EPnP: An Accurate O(n) Solution to the PnP Problem. *Int. J. Comput. Vis.*, 81:155–166, 2009.