Human Interaction and Emerging Technologies (IHIET 2026), Vol. 236, 2026, 522–530
https://doi.org/10.54941/ahfe1008127

Object Perception Pipeline for Industrial Disassembly: From 6D Pose to 3D Localization in Automotive Robotics

Evangelos Sartinas, Athina Zacharia, and Maria Pateraki

National Technical University of Athens, School of Rural, Surveying and Geoinformatics Engineering, Athens, 15772, Greece

ABSTRACT

Industrial disassembly processes require robust and efficient perception systems capable of handling heterogeneous objects under real-world constraints. In automotive manufacturing and disassembly scenarios, components exhibit significant variability in geometry, size, symmetry, and placement, making it impractical to rely on a single, uniform object pose estimation strategy. At the same time, such environments impose strict requirements on the perception systems in terms of reliability, computational efficiency, and ease of deployment. This work presents an instance-based object perception pipeline with task-driven object pose estimation, developed within the SOPRANO EU project, for automated automotive door disassembly. The pipeline assumes a known set of object instances and depending on the objects' geometric characteristics, the manipulation task requirements, the system performs (i) full 6D pose estimation, (ii) planar-constrained 3D localization via RGB-D lifting, or (iii) planar localization with structured multi-instance refinement and instance identification. The perception process consists of two stages: pose formulation selection and execution of the corresponding estimation method. Model-based 6D pose estimation using RGB data is applied to rigid objects, while RGB-D-based lifting of 2D detections is used for planar or quasi-planar elements. For structured arrangements of multiple instances of a single object type, such as screw arrays, a multi-instance matching strategy ensures consistent indexing and reduces ambiguity. The system is deployed as a modular perception service and validated in an automotive disassembly pilot. Experimental results demonstrate high accuracy across heterogeneous tasks, highlighting the benefits of aligning perception outputs with task-specific requirements.

Keywords: 6D pose estimation, Instance-based perception, Task-driven localization, RGB-D perception, Industrial robotics, Disassembly automation

INTRODUCTION

Industrial disassembly processes require robust perception systems capable of identifying and localizing heterogeneous objects under real-world operational constraints. In automotive environments, components vary significantly in geometry, size, symmetry, and spatial configuration, making it impractical to rely on a single pose estimation strategy. At the same time, production settings demand solutions that are accurate, reliable, and computationally efficient.

Received April 24, 2026; Revised June 26, 2026; Accepted July 15, 2026; Available online August 25, 2026

Most existing approaches treat object perception as a uniform problem, typically formulating it as full six-degree-of-freedom (6D) pose estimation regardless of object characteristics or task requirements (Thalhammer et al., 2024; Guan et al., 2024). While effective for rigid 3D objects, this formulation can introduce unnecessary complexity and may lead to ambiguous or unstable predictions, particularly for planar, symmetric, or weakly textured objects commonly encountered in industrial scenarios (Manhardt et al., 2019; Hsiao et al., 2024; Hodan et al., 2017).

In this work, we propose a task-driven, instance-based perception pipeline that formulates the perception problem by solving only the degrees of freedom required for each object and task. Rather than estimating the full 6D pose for all objects, the system assigns each object type the minimal set of degrees of freedom required for reliable manipulation, based on its geometric properties and functional context (e.g., grasping, unscrewing, placement).

The proposed pipeline is developed within the SOPRANO EU project and deployed in an automotive door disassembly scenario. It supports multiple levels of geometric inference, ranging from full 6D pose estimation for rigid components to constrained 3D localization for planar objects and structured multi-instance localization for repeated elements. By estimating only the necessary degrees of freedom per object, the system avoids common failure modes of over-generalized approaches while reducing computational complexity and dependence on extensive annotated training data.

This task-driven formulation enables efficient and scalable deployment, as full 6D estimation, and its associated requirement for CAD models and dense supervision, is only applied where strictly necessary. Simpler objects are handled using lightweight detection-based methods combined with geometric reasoning. As a result, aligning perception outputs with task-specific requirements provides a practical and robust solution for real-world industrial robotics.

RELATED WORK

6D object pose estimation aims to recover the rigid transformation between the object and camera coordinate systems, typically represented as a translation and rotation in 3D. State-of-the-art methods are extensively reviewed in several surveys (Clément et al., 2019; Şahin et al., 2018; Wu et al., 2020; Zhu et al., 2022) which categorize approaches based on prior information, sensor modality, number of views, and inference strategy. In recent years, deep learning-based methods have become dominant, particularly for estimating pose from single RGB or RGB-D images (Hodan et al., 2024).

Instance-based methods such as EPOS (Hodan et al., 2020), ZebraPose (Su et al., 2022) and Dope (Tremblay et al., 2018) achieve high accuracy by leveraging CAD models and supervised training to recover full 6D pose. These approaches are particularly effective when accurate object models and sufficient annotated data are available, and they consistently achieve strong performance on benchmark datasets. More recently, generalizable approaches such as FoundationPose (Wen et al., 2024), Freezev2 (Caraffa et al., 2024) and OnePose++ (He et al., 2022) have been proposed to handle novel objects without object-specific training, reducing reliance on annotated datasets and enabling more flexible deployment. However, several challenges persist in industrial environments, including object symmetries, reflective surfaces, occlusions, and weak textures, which can lead to ambiguous or unstable pose predictions.

Complementary to 6D pose estimation, 2D object detection methods such as YOLO (Redmon et al., 2016) and Mask R-CNN (He et al., 2017) are widely used for object detection and instance segmentation, and support object localization when combined with depth data to recover 3D positions. These approaches are computationally efficient and well-suited for real-time applications, making them attractive for industrial deployment. Semantic segmentation methods further enhance localization by providing pixel-level information, enabling more precise spatial reasoning. Additionally, training-free segmentation approaches such as CNOS (Nguyen et al., 2023) enable object mask extraction without requiring object-specific annotations, making them suitable for novel-object scenarios.

However, 2D detection-based approaches do not recover full object pose and are therefore limited in tasks requiring accurate orientation estimation. Additionally, their performance may degrade in scenarios involving small objects, occlusions, or visual ambiguity. While depth-based lifting partially addresses this limitation by providing 3D position, it does not fully resolve ambiguities in orientation or object identity in structured environments.

Applying full 6D pose estimation to all objects can introduce unnecessary complexity and lead to unstable predictions in cases of symmetric or planar objects, while relying solely on 2D detection limits applicability in tasks requiring precise orientation. Most existing approaches treat pose estimation as a uniform problem, without considering task-specific requirements or object geometry. In contrast, this work proposes a task-driven, instance-based perception pipeline that dynamically selects the minimal pose representation required for each object. This constraint-driven formulation improves robustness and computational efficiency, making it well-suited for industrial disassembly scenarios.

USE CASE: AUTOMOTIVE DOOR DISASSEMBLY

The deployed scenario involves the automated disassembly of vehicle doors for remanufacturing and recycling. Robotic systems operate on a fixed and known set of components, including the door panel, screws, internal modules, and tools placed on the door fixture and nearby worktables. As the object set is predefined, the perception problem is inherently instance-based rather than category-based. The robotic system consists of a collaborative manipulator (UR10) used for fine manipulation tasks such as screw detection and unscrewing, and a mobile robotic platform equipped with a manipulator for handling tools and larger components. The perception pipeline provides task-specific outputs to both platforms, supporting precise grasping, alignment, and interaction within the disassembly process.

The objects exhibit significant geometric diversity, ranging from rigid 3D components (e.g., door panel, tools, hinges) to small quasi-planar elements (e.g., screws, clips), as well as symmetric or rotationally ambiguous objects. In addition, several elements are arranged in structured configurations, such as repeated planar objects with known spatial layouts (e.g., screw arrays), while others may be partially occluded within the door structure.

Figure 1 illustrates representative examples of these perception tasks, including (a) objects requiring full 6D pose estimation and (b) screws arranged in predefined configurations on the door frame.

These characteristics impose different perception requirements depending on the manipulation task. Rigid components such as tools or the door panel require full 6D pose estimation for accurate grasping and interaction. In contrast, planar or quasi-planar objects typically require only 3D position with constrained orientation, as their pose is partially defined by the supporting surface. For multiple instances of the same object type arranged in known configurations, such as screw arrays, the task reduces to localizing each instance and assigning it to its expected position.

This variability motivates a task-driven perception approach in which the required geometric representation is defined for each object instance based on its geometry and manipulation role. The proposed pipeline follows this formulation by selecting, at design time, the appropriate pose representation and corresponding estimation method for each object. As a result, it outputs only the necessary geometric information required for the task, reducing unnecessary complexity while improving robustness in structured industrial environments.

(a) (b)

Figure 1: Indicative examples of the 6D object pose and 3D localization tasks: (a) components requiring full 6D pose estimation and (b) screws in predefined configurations on the door frame.

PIPELINE OVERVIEW

The proposed perception pipeline follows a task-driven design in which each object instance is associated, at system configuration time, with a predefined pose representation, and assigned to the appropriate perception module (e.g., 6D pose estimation or detection-based localization) responsible for estimating that representation. Rather than applying a uniform pose estimation strategy, the pipeline selects the appropriate representation (i.e., the degrees of freedom to be estimated) and corresponding estimation method for each object.

The pipeline operates in two stages: (1) representation selection, where the required degrees of freedom are defined for each object, and (2) estimation, where the corresponding perception method is executed. The considered representations range from full 6D pose estimation for rigid components to constrained 3D localization for planar or quasi-planar objects, as well as structured localization for repeated objects arranged in known configurations.

In the case of rigid 3D objects such as tools or large assemblies, full 6D pose estimation is performed using model-based methods (e.g., EPOS, ZebraPose) operating on RGB input and CAD models. For planar or quasi-planar objects, 2D detections (e.g., YOLO, Mask R-CNN) are lifted into 3D using depth data, while geometric constraints are applied to restrict irrelevant degrees of freedom. For repeated objects arranged in structured layouts, such as screw arrays, detections are further refined through geometric matching and instance-level correspondence to predefined spatial configurations.

Figure 2 illustrates the overall pipeline. The system recovers only the degrees of freedom required by the manipulation task and the remaining components are derived from geometric or environmental priors, such as support surface orientation, object symmetries, or known spatial layouts. For example, in the case of screws mounted on a door panel, only the 3D position is estimated from detection, while the orientation is inferred from the known pose of the supporting surface. Similarly, for structured configurations of repeated objects, detections are aligned with expected layouts to ensure consistent indexing.

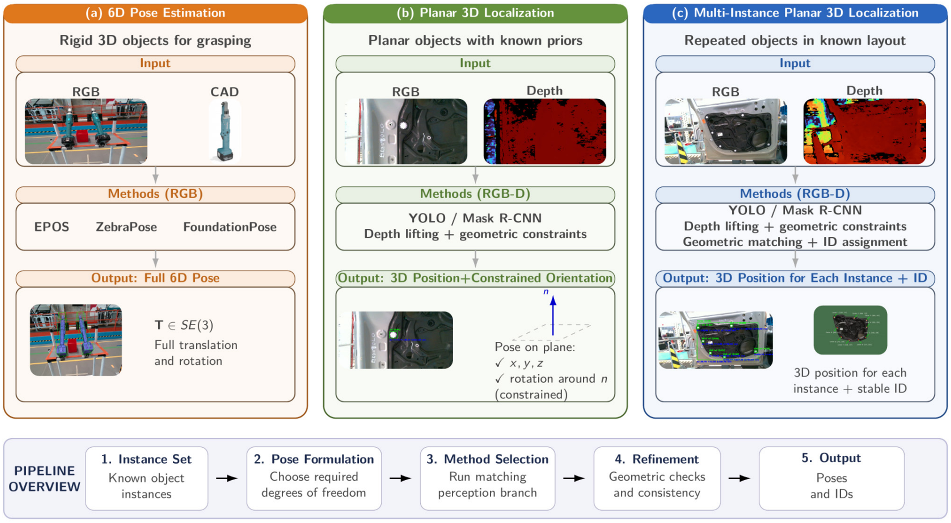

Figure 2: Overview of the proposed instance-based perception pipeline. For each object, a task-specific pose formulation is selected, enabling (a) full 6D pose estimation, (b) constrained 3D localization via RGB-D lifting, or (c) structured multi-instance matching. By estimating only task-relevant degrees of freedom and incorporating geometric priors, the pipeline improves robustness and efficiency while producing consistent pose outputs for manipulation.

EXPERIMENTAL VALIDATION AND OBSERVATIONS

The perception pipeline was evaluated on representative tasks within the automotive disassembly scenario, including (i) 6D pose estimation for tool grasping and (ii) planar localization for screw detection and unscrewing.

6D POSE ESTIMATION

The evaluation setup follows the protocol and benchmarks introduced in the ***IndustryShapes*** (Sapoutzoglou et al., 2026), a recent RGB-D dataset designed specifically for industrial 6D pose estimation. *IndustryShapes* provides a realistic testbed for evaluating perception methods under conditions that closely resemble deployment environments, including cluttered scenes, multiple object instances, occlusions, and challenging object properties such as weak texture, symmetries, and reflective surfaces.

The dataset consists of two complementary subsets: (i) the ***classic set***, used for instance-level evaluation with annotated 6D poses, and (ii) the ***extended set***, designed for evaluating generalizable and novel-object methods, including RGB-D onboarding sequences. The data were captured in realistic industrial environments using RGB-D sensors, and include both controlled single-object training scenes and complex multi-object test scenarios, enabling evaluation under both ideal and deployment-like conditions.

To reflect the different estimation paradigms relevant to the proposed pipeline, we evaluate one representative instance-based method (EPOS) and a method capable of handling novel objects (FoundationPose). This selection allows us to analyze the trade-offs between object-specific training and generalization in realistic industrial scenarios.

For the experiments, instance-based methods were trained on the classic training split, while generalizable methods were evaluated without object-specific retraining, following the dataset protocol. For FoundationPose, object masks were obtained using CNOS predictions instead of ground-truth masks or trained detectors (e.g., YOLO), as CNOS does not require object-specific training, thereby preserving the truly novel-object setting while better reflecting real-world conditions, where segmentation must be provided by an upstream perception module. Performance was assessed using standard 6D pose metrics, including ADD, MSPD, MSSD, and VSD, following the BOP evaluation protocol (Hodan et al., 2020).

The performance of instance-based and generalizable pose estimation methods was evaluated using EPOS and FoundationPose. Both methods achieved high accuracy, satisfying the requirements for stable robotic grasping. FoundationPose, despite operating without object-specific training, demonstrated performance comparable to or slightly exceeding EPOS in some metrics, highlighting the trade-off between specialization and generalization.

These results support the need for adaptive pose formulation, as different methods are better suited to different object and task conditions.

Table 1: 6D pose estimation performance.

Method	ADD ↑	MSPD ↑	MSSD ↑	VSD ↑	Average Recall (AR) ↑
EPOS	0.64	0.54	0.56	0.41	0.51
FoundationPose	0.59	0.55	0.56	0.42	0.51

PLANAR LOCALIZATION AND SCREW DETECTION

For screw localization, a dedicated detection and refinement pipeline was employed, combining synthetic-data-trained detectors with geometric constraints and structured spatial matching.

To support this task, a dedicated RGB-D screw dataset was constructed using data captured in the CRF industrial environment. The dataset includes manually annotated screw instances under realistic conditions, such as varying illumination, occlusions, and viewpoints. In addition, synthetic training data were generated using CAD models and photorealistic rendering to increase variability and improve robustness to challenging visual properties, including reflections and weak texture.

A YOLO-based detector was trained on the combined dataset and deployed in a two-stage operational setup. In **Stage 1**, multiple screws are visible in array-like configurations, and the task consists of detecting all instances and associating them with their expected spatial layout. In **Stage 2**, after closer robot placement, only a single screw is observed, requiring highly precise localization for tool alignment.

In both stages, 2D detections are lifted to 3D using depth information and refined using geometric priors. For structured configurations, detections are additionally matched to predefined layouts to ensure consistent instance identification.

Experimental results demonstrate high performance across operating conditions, with detection accuracy reaching up to 99.7% mAP at IoU 0.5. Under stricter localization requirements in Stage 2 (IoU 0.9), performance remained consistently high, confirming suitability for precise robotic unscrewing tasks.

The integration of geometric constraints and structured multi-instance matching further improves robustness, reducing ambiguity among visually identical elements and ensuring reliable localization across both operational stages.

Table 2: Screw detection performance (planar localization).

Stage	IoU Threshold	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)	Task-Specific Accuracy (%)
Stage 1 (multi-screw)	65%	99.4	61.3	99.7	97.1	96.6
Stage 2 (single-screw)	90%	99.7	89.1	99.7	100.0	92.4

CONCLUSION

This work presented a task-driven, instance-based perception pipeline for industrial robotic manipulation, motivated by the specific requirements of automotive door disassembly. Unlike conventional approaches that treat object perception as a uniform problem, the proposed framework explicitly incorporates object geometry and task constraints into the design of the perception system. By assigning a suitable pose representation to each object instance at design time, the pipeline enables the use of tailored estimation strategies, ranging from full 6D pose estimation to constrained 3D localization and structured multi-instance matching.

Experimental validation on the IndustryShapes dataset and on a dedicated screw detection dataset collected in a real industrial environment demonstrates the effectiveness of this approach. Instance-based methods such as EPOS achieve high accuracy for tasks requiring precise pose estimation, while generalizable methods such as FoundationPose provide flexibility for handling novel objects without retraining. In parallel, the proposed planar localization pipeline achieves high detection accuracy and reliable instance indexing for repetitive structures, supporting precise robotic operations such as unscrewing.

The results highlight that adapting the pose representation to the task requirements improves robustness and reduces unnecessary estimation complexity, particularly in challenging scenarios involving symmetries, occlusions, and weak textures. This confirms the benefits of a constraint-driven formulation over uniform perception pipelines in structured industrial settings.

Future work will focus on extending the approach to more dynamic environments, improving robustness under severe occlusions and reflective materials, and further integrating generalizable pose estimation methods. Additionally, automating parts of the design-time formulation process and expanding dataset coverage will support increased scalability and broader applicability across industrial domains.

ACKNOWLEDGMENT

This work was funded by the HEU programme SOPRANO (GA No 101120990).

REFERENCES

Clement, F., Shah, K., & Pancholi, D. (2019). A Review of methods for Textureless Object Recognition. arXiv preprint arXiv:1910.14255.

Guan, J., Hao, Y., Wu, Q., Li, S., & Fang, Y. (2024). A survey of 6dof object pose estimation methods for different application scenarios. Sensors, 24(4), 1076.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961–2969).

He, X., Sun, J., Wang, Y., Huang, D., Bao, H., & Zhou, X. (2022). Onepose++: Keypoint-free one-shot object pose estimation without CAD models. Advances in Neural Information Processing Systems, 35, 35103–35115.

Hodan, T., Barath, D., & Matas, J. (2020). Epos: Estimating 6d pose of objects with symmetries. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 11703–11712).

Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., & Zabulis, X. (2017, March). 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) (pp. 880–888). IEEE.

Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V. N., Wang, G., Brachmann, E., ... & Matas, J. (2024). Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5610–5619).

Hsiao, T. C., Chen, H. W., Yang, H. K., & Lee, C. Y. (2024). Confronting ambiguity in 6d object pose estimation via score-based diffusion on se (3). In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 352–362).

Manhardt, F., Arroyo, D. M., Rupprecht, C., Busam, B., Birdal, T., Navab, N., & Tombari, F. (2019). Explaining the ambiguity of object detection and 6d pose from visual data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6841–6850).

Nguyen, V. N., Groueix, T., Ponimatkin, G., Lepetit, V., & Hodan, T. (2023). Cnos: A strong baseline for cad-based novel object segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 2134–2140).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779–788).

Sahin, C., & Kim, T. K. (2018). Recovering 6D object pose: A review and multi-modal analysis. In Proceedings of the European conference on computer vision (ECCV) workshops (pp. 0–0).

Sapoutzoglou, P., Vaggelis, O., Zacharia, A., Sartinas, E., and Pateraki, M. (2026). IndustryShapes: An RGB-D benchmark dataset for 6D object pose estimation of industrial assembly components and tools. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2026), Vienna, Austria.

Su, Y., Saleh, M., Fetzer, T., Rambach, J., Navab, N., Busam, B., ... & Tombari, F. (2022). Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6738–6748).

Thalhammer, S., Bauer, D., Hönig, P., Weibel, J. B., Garcia-Rodriguez, J., & Vincze, M. (2024). Challenges for monocular 6-d object pose estimation in robotics. IEEE Transactions on Robotics, 40, 4065–4084.

Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., & Birchfield, S. (2018). Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790.

Wen, B., Yang, W., Kautz, J., & Birchfield, S. (2024). Foundationpose: Unified 6d pose estimation and tracking of novel objects. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 17868–17879).

Wu, J., Yin, D., Chen, J., Wu, Y., Si, H., & Lin, K. (2020, April). A survey on monocular 3D object detection algorithms based on deep learning. In Journal of Physics: Conference Series (Vol. 1518, No. 1, p. 012049). IOP Publishing.

Zhu, Y., Li, M., Yao, W., & Chen, C. (2022, June). A review of 6d object pose estimation. In 2022 IEEE 10th joint international information technology and artificial intelligence conference (ITAIC) (Vol. 10, pp. 1647–1655). IEEE.