Improving 6D Object Pose Estimation via Clustering-based Multi-View Consensus

Bálint Áron Janik[1] and Viktor Varga[1]

Eötvös Loránd University, Budapest, Hungary[1]
f5z688@inf.elte.hu, vv@inf.elte.hu

Abstract. Most 6D object pose estimation pipelines rely heavily on 2D detections, which can be inaccurate in cluttered and ambiguous scenes. To overcome this issue, we propose a lightweight framework that improves upon 2D object detections using multi-view consensus. We formulate cross-view instance association as an agglomerative clustering process that groups detection candidates based on geometric consistency. By leveraging multiple views, the approach recovers details missed in single-view predictions, improving detection quality and pose estimation accuracy. As a lightweight pre-processing step, it introduces minimal computational overhead and can be seamlessly integrated into existing pipelines. We release the source code.

1 Introduction

6D object pose estimation aims to recover the 3D rotation and translation of an object relative to the camera [3,10]. It is a central problem in computer vision with applications in robotic manipulation [13], augmented reality [1], and digital twin systems [7].

Recent single-view methods rely on feature matching between image observations and object models, enabling support for novel objects with only a CAD model or reference images [9,12]. However, reliance on a single image makes them susceptible to occlusion and ambiguity, particularly in cluttered scenes with textureless objects.

Multi-view approaches address these limitations by leveraging multiple observations. CosyPose aggregates and jointly refines pose hypotheses across views, improving robustness through geometric consistency [5], while other works similarly exploit multi-view consistency [2,6]. However, existing multi-view methods typically operate after pose hypotheses have already been generated. As a result, their performance remains constrained by the quality of the initial 2D candidates: missed detections remove visible objects from consideration, while false positives can introduce inconsistent hypotheses that increase ambiguity during downstream pose estimation. Furthermore, progress in open-source multi-view 6D pose estimation has been limited, with CosyPose being one of the latest widely available systems [5], while most recent state-of-the-art methods focus on single-view inference.

6D pose estimation pipelines typically follow a two-stage design, where image-based object detection generates candidate regions for subsequent pose estimation. Overall performance depends heavily on the quality of these regions. Although recent category-agnostic detectors such as CAD-based Novel Object Segmentation (CNOS) improve generalization, they remain unreliable under occlusion and ambiguity [8]. We observe that the majority of erroneous pose estimations are a result of inaccurate object detections, including missed detections of visible objects and false-positive detections assigned to objects that are absent.

This motivates using multi-view geometry earlier in the pipeline, not only to refine pose hypotheses but also to validate and improve the 2D detections on which those hypotheses depend. To study this gap, we focus on the T-LESS dataset from the BOP benchmark, which features textureless objects and provides accurate camera pose annotations, making it well suited for multi-view reasoning under occlusion [4].

In this paper, we present a lightweight multi-view consensus framework for refining 2D object detections. We model cross-view object association as an agglomerative clustering problem, grouping candidates by geometric consistency. This suppresses spurious detections and reinforces consistent observations across views, resulting in more robust detections and reliable cross-view correspondences. We demonstrate that refining detections before pose estimation improves downstream RGB-D 6D pose estimation results while adding minimal computational overhead. The source code is available at: `https://github.com/balintjanik/icpr26_cesi_workshop_6dpose`.

2 Method

The proposed method is designed as a plug-in refinement stage for existing RGB-D 6D object pose estimation methods. Given a set of views, an off-the-shelf detector first produces per-view object masks and confidence scores. Our method refines these detections by enforcing multi-view geometric consistency before passing the resulting masks to an unchanged downstream pose estimator.

Given calibrated RGB-D images and per-image object detections, we first apply non-maximum suppression on the detections (with an IoU threshold of η). Then, each detection is lifted to 3D by back-projecting pixels within its mask into a common world coordinate system with depth information provided by the camera. Given a pixel $u = (x, y)$ with depth d, its 3D position in camera coordinates is computed as

$$p_c = d\, K^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \tag{1}$$

and transformed to world coordinates as $p_w = R^T(p_c - t)$, where K denotes the camera intrinsics and (R, t) the extrinsics. For each detection i, we obtain a set of 3D points X_i with centroid c_i, confidence score s_i (estimated by the detector), view index v_i and mask m_i.

2.1 Multi-view clustering

We associate detections across views via agglomerative clustering. The pairwise distance between detections i and j is defined as

$$d_{ij} = \begin{cases} \dfrac{\|c_i - c_j\|_2}{s_i s_j + \varepsilon}, & v_i \neq v_j, \\ \infty, & v_i = v_j, \end{cases} \tag{2}$$

where ε is a small constant added for numerical stability. This encourages the grouping of geometrically consistent detections, favoring high-confidence matches. Assigning infinite distance to detections from the same view prevents intra-view clustering. Clustering is performed using average linkage with a distance threshold $\tau_o = \alpha \cdot D_o$, where D_o is the object diameter and α is a fixed scaling factor. This makes the threshold adaptive to object scale.

After clustering, each cluster is restricted to at most one detection per view by retaining, for each view, the highest confidence detection. Let $V_{\mathcal{C}} = \{v_i \mid i \in \mathcal{C}\}$ denote the set of supporting views. Clusters are ranked by $|V_{\mathcal{C}}|$ in descending order, with ties broken by the cumulative detection confidence $\sum_{i \in \mathcal{C}} s_i$.

2.2 Consensus mask reconstruction

For each retained cluster $\mathcal{C}$, we define the aggregated 3D point set $P_{\mathcal{C}} = \bigcup_{i \in \mathcal{C}} X_i$. For each view v, the cluster mask $M_{\mathcal{C}}^{(v)}$ is defined as

$$M_{\mathcal{C}}^{(v)} = \begin{cases} m_i, & \text{if } \exists i \in \mathcal{C} \text{ with } v_i = v, \\ T_{\text{morph}}\left(\hat{M}_{\mathcal{C}}^{(v)}\right), & \textit{otherwise}, \end{cases} \tag{3}$$

where $\hat{M}_{\mathcal{C}}^{(v)}$ is obtained by projecting $P_{\mathcal{C}}$ into view v via $u \sim K(Rp_w + t)$ for $p_w \in P_{\mathcal{C}}$, followed by rasterization into a binary mask, and T_{morph} denotes a morphological closing operation. This procedure reconstructs masks in views where detections are absent or unreliable.

3 Experiments

We evaluate the proposed multi-view consensus framework on the BOP'19 evaluation subset of the T-LESS dataset, featuring textureless objects under heavy occlusion and clutter. With our method included in the pipeline, we will first measure the improvement in detection quality, and then assess the impact on final pose estimation results. Pose estimation is performed using the FoundationPose pipeline. We experiment with two state-of-the-art alternatives for object localization in the pipeline: (i) segmentation masks from CNOS (a category-agnostic detector); and (ii) bounding-box detections from the Pix2Pose detector (RetinaNet) [11], which is trained in an object-specific manner.

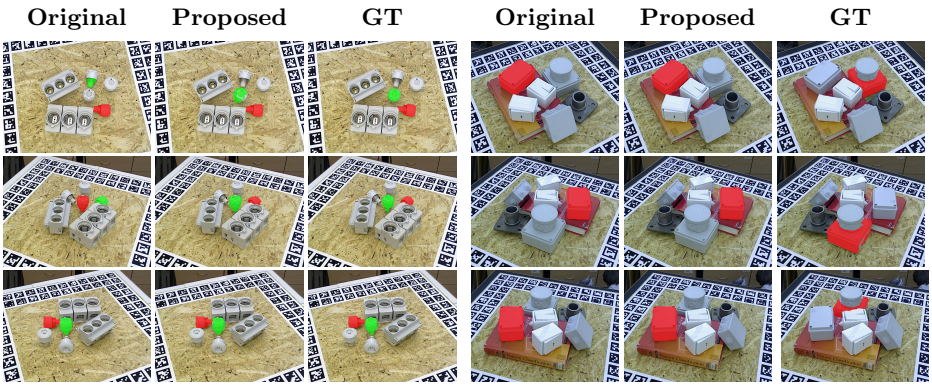

Fig. 1: Examples of original and corrected object proposals from the T-LESS dataset. The rows show different views. The first three columns show a successful correction, while the other three show a failure case.

We compare our multi-view consensus approach, which clusters, filters and reconstructs detections across views, with a single-view baseline where the top-K detections are selected independently per image based on confidence. K is fixed to the known number of object instances per scene. Optimal hyperparameters values $\alpha = 2.0$, $\eta = 0.7$ were found empirically.

3.1 Detection reliability improvements

We evaluate segmentation quality using Intersection-over-Union (IoU) with ground-truth masks (Table 1). Results for 1 view corresponds to the original pipeline without our method using the CNOS detector. Increasing the number of views improves mean IoU, whereas median IoU shows only minor gains. This is caused by an improvement in robustness rather than per-instance accuracy, as the ratio of total detection failures has been significantly reduced. Figure 1 shows qualitative examples of refined detections. In most cases, our multi-view consensus improves robustness by restoring detections missed in individual views. However, the last three columns show a limitation: if an object is not detected in any view, our method cannot recover it.

The proposed module introduces minimal computational overhead, adding just 11 milliseconds to the time taken for each view on T-LESS. Runtime was measured on a simple desktop machine featuring an Intel Xeon CPU E5-2620 v4 (2.10GHz) CPU and a single Nvidia Titan RTX GPU.

3.2 Pose estimation improvements

We evaluate the impact of the refined detections on 6D pose estimation using the FoundationPose pipeline with both detector alternatives. Importantly, the pose estimation procedure itself remains single-view. Multi-view information is used

Table 1: Segmentation performance as a function of the number of views on the BOP'19 subset of the T-LESS dataset.

Views	Mean IoU	Median IoU
1	0.485	0.742
3	0.541	0.751
5	0.561	0.753
10	**0.589**	**0.769**

Table 2: Pose estimation performance of the proposed method across varying numbers of views on the BOP'19 subset of the T-LESS dataset. Multi-view information is used to refine object masks estimated by CNOS (left) and bounding boxes given by Pix2Pose RetinaNet (right). Refined detections are fed into the FoundationPose pipeline in both cases. The BOP'19 standard metrics are used.

	CNOS					Pix2Pose (RetinaNet)				
Views	1	2	3	5	10	1	2	3	5	10
AR	0.579	0.653	0.678	0.690	**0.704**	0.577	0.673	0.698	**0.709**	0.707
MSPD	0.601	0.679	0.705	0.717	**0.732**	0.599	0.700	0.727	**0.738**	0.736
MSSD	0.590	0.666	0.691	0.702	**0.717**	0.590	0.686	0.711	**0.722**	0.720
VSD	0.545	0.614	0.637	0.650	**0.664**	0.543	0.634	0.657	**0.668**	0.665

only to refine 2D detections prior to pose estimation. Table 2 reports performance as a function of the number of views used for detection refinement. Using only five views with the CNOS detector increases the average recall by 11.1 percentage points, according to the standard BOP'19 benchmark recall metric.

Performance improves with more views, but the improvement becomes less significant after several views. This indicates that most recoverable failures are already resolved with a small number of observations. For Pix2Pose, performance saturates around five views and slightly decreases when ten views are used, likely because additional views may also introduce noisy bounding boxes or false-positive detections. These detections can occasionally reinforce incorrect candidates. Overall, the results show that improving detection reliability directly translates into more accurate 6D pose estimates.

4 Conclusion

The results suggest that many pose estimation errors stem from inaccurate 2D detections rather than the pose estimator itself. By enforcing cross-view consistency, our plug-in refinement stage reduces detection errors and improves pose accuracy. Its reliance on known extrinsics limits uncalibrated use, and remain-

ing failures show that geometric aggregation cannot resolve systematic errors or persistent ambiguity.

Acknowledgments. The project was carried out as part of the TÉT project (No. 2024-1.2.5-TÉT-2024-00019) funded by the National Research, Development and Innovation Office.

References

1. Aguirrezabal, P., Aguinaga, I., Alvarez-Gila, A.: Monocular rgb 6d object pose estimation for augmented reality: a survey. Virtual Reality **30** (02 2026). https://doi.org/10.1007/s10055-026-01315-4
2. Duffhauss, F., Demmler, T., Neumann, G.: Mv6d: Multi-view 6d pose estimation on rgb-d frames using a deep point-wise voting network. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3568–3575 (2022). https://doi.org/10.1109/IROS47612.2022.9982268
3. Guan, J., Hao, Y., Wu, Q., Li, S., Fang, Y.: A survey of 6dof object pose estimation methods for different application scenarios. Sensors **24**(4) (2024). https://doi.org/10.3390/s24041076, https://www.mdpi.com/1424-8220/24/4/1076
4. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: 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 (2017)
5. Labbé, Y., Carpentier, J., Aubry, M., Sivic, J.: Cosypose: Consistent multi-view multi-object 6d pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 574–591. Springer International Publishing, Cham (2020)
6. Li, A., Schoellig, A.P.: Multi-view keypoints for reliable 6d object pose estimation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 6988–6994 (2023). https://doi.org/10.1109/ICRA48891.2023.10160354
7. Nguyen, K., Dong, Q., Pham, T., Do-Duy, T., Nguyen, K., Tran, C.C., Tran, H., Nguyen, Q.C.: Cad-guided 6d pose estimation with deep learning in digital twin for industrial collaborative robot manipulation. EAI Endorsed Transactions on AI and Robotics **4** (10 2025). https://doi.org/10.4108/airo.9676
8. Nguyen, V.N., Groueix, T., Ponimatkin, G., Lepetit, V., Hodan, T.: Cnos: A strong baseline for cad-based novel object segmentation. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 2126–2132 (2023). https://doi.org/10.1109/ICCVW60793.2023.00227
9. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: Gigapose: Fast and robust novel object pose estimation via one correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9903–9913 (June 2024)
10. Nguyen, V.N., Tyree, S., Guo, A., Fourmy, M., Gouda, A., Lee, T., Moon, S., Son, H., Ranftl, L., Tremblay, J., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Birchfield, S., Matas, J., Labbe, Y., Sundermeyer, M., Hodan, T.: Bop challenge 2024 on model-based and model-free 6d object pose estimation (2025), https://arxiv.org/abs/2504.02812
11. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: 2019 IEEE/CVF International Conference

on Computer Vision (ICCV). pp. 7667–7676 (2019). https://doi.org/10.1109/ICCV.2019.00776
12. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17868–17879 (June 2024)
13. Xia, W., Zheng, H., Xu, W., Xu, X.: Large vision-language models enabled novel objects 6d pose estimation for human-robot collaboration. Robotics and Computer-Integrated Manufacturing **95**, 103030 (2025). https://doi.org/https://doi.org/10.1016/j.rcim.2025.103030, https://www.sciencedirect.com/science/article/pii/S0736584525000845