POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Zero-shot Model-free 6D Object Pose Estimation in RGB-D Videos

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Simone Paloschi

Advisor: Prof. Federica Arrigoni

Co-advisor: Angelo Iacoella, Giacomo Boracchi

Academic year: 2025-2026

1. Introduction

The 6D pose of a rigid object encodes its position and orientation in three-dimensional space, combining three translational and three rotational degrees of freedom. Estimating this pose across the frames of a video, known as pose tracking, is a fundamental task in robotics and augmented reality. Classical approaches rely on a pre-built 3D model of the object, and even recent model-free methods require a dedicated set of reference images with known poses before tracking can begin. In many real-world scenarios, however, neither is practical to obtain, motivating a growing class of zero-shot model-free methods that require no geometric prior and no reference observations.

This thesis explores this setting on RGB-D videos, where each frame pairs a standard color image with a per-pixel depth map recording the distance from the camera. As tracking proceeds, the system naturally accumulates posed observations of the object from multiple viewpoints, progressively covering its surface and creating an opportunity to improve the 3D object representation over time. We present a novel pipeline that exploits recent advances in image-to-3D methods to generate a 3D mesh of the object from the first frame, and progressively refines it using the color and depth information accumulated during tracking. We evaluate the approach on a challenging benchmark with hand-manipulated objects and show that zero-shot model-free tracking is viable at a small accuracy cost relative to model-based methods, and that mesh refinement consistently improves both geometry and pose accuracy, opening practical possibilities in settings where preparing a 3D model or a reference collection is too costly. The main contributions of this thesis are:

- A fully zero-shot model-free pipeline for 6D pose tracking from a single RGB-D video.
- An empirical study of the current capabilities and limitations of zero-shot model-free pose tracking, showing how close this setting can get to model-based performance.
- An evaluation on a challenging video dataset comparing the proposed method against state-of-the-art approaches.

2. Problem Formulation

The problem addressed is 6D pose tracking and mesh completion for unknown objects, with no prior knowledge of their geometry or appearance. The input is an RGB-D video sequence, the camera calibration matrix $\mathbf{K}$, and a segmentation prompt in the first frame identifying the

object of interest, such as two points clicked on the object. The two outputs are defined as follows.

6D Pose Sequence. For each frame i, the system estimates a pose $\{\mathbf{R}_i, \mathbf{t}_i\}$, see Figure 1, describing the object's position and orientation relative to the camera, where $\mathbf{R}_i \in \mathrm{SO}(3)$ is a rotation matrix and $\mathbf{t}_i \in \mathbb{R}^3$ is a translation vector. These are combined into a single 4×4 transformation matrix:

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

which maps any point from the object's local coordinate system into the camera frame.

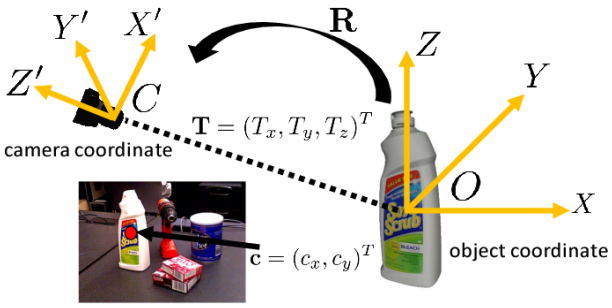

Figure 1: 6D pose. The rotation $\mathbf{R}$ and translation $\mathbf{T} = (T_x, T_y, T_z)^\top$ map the object's local coordinate frame $(O; X, Y, Z)$ into the camera frame $(C; X', Y', Z')$. $\mathbf{c} = (c_x, c_y)^\top$ is the camera origin C projected into image-space.

Reconstructed 3D Mesh. The system also produces a mesh $\mathcal{M}$ representing the complete surface geometry of the object, including regions never directly observed. A mesh consists of a set of vertices $\mathbf{V} = \{\mathbf{v}_i \in \mathbb{R}^3\}$, a set of triangular faces connecting them, and per-vertex or per-face appearance information encoding color and texture. The mesh is defined in the object's local coordinate system, and the estimated poses are used to place it in the camera frame at each frame.

Assumptions. The tracked object is assumed to be rigid. Its motion between consecutive frames is assumed to be smooth. Only one object of interest is present in the scene.

3. Related Works

3.1. Object Segmentation

Before pose estimation can begin, the target object must be identified. The standard approach is to use an off-the-shelf segmentation method, which produces a pixel-level mask indicating which pixels belong to the object of interest. Mesh completion additionally requires the mask to be propagated across all subsequent frames, to be able to filter the object information. Video Object Segmentation methods address this by leveraging temporal coherence, maintaining consistent masks over time given only a first-frame annotation.

3.2. 6D Object Pose Estimation

Traditional 6D pose estimation relied on handcrafted features and explicit 3D models, which limited generalization. Deep learning progressively relaxed these constraints, replacing manual features with learned representations that operate from either a 3D mesh or a set of reference images. FoundationPose [4] represents the current state-of-the-art along this direction: it generates multiple pose hypotheses, refines each one by rendering the mesh and comparing it with the observed image through a dedicated network, and selects the best result via a scoring network. The method supports both a full estimation mode for the first frame and an efficient tracking mode for subsequent frames that simply refines the previous estimate.

3.3. Joint Pose Estimation and Reconstruction

Early methods for jointly estimating pose and reconstructing object geometry were adaptations of SLAM, which incrementally builds a 3D map during camera motion. More recent approaches encode object geometry as a Signed Distance Field (SDF) - a neural network that, given any 3D point, predicts its distance to the nearest object surface, implicitly defining the object's shape as a continuous function learnable from a set of images. However, these methods faithfully represent only the regions they have observed, leaving unobserved parts permanently incomplete, and they require extensive object coverage before the geometry is reliable enough for tracking.

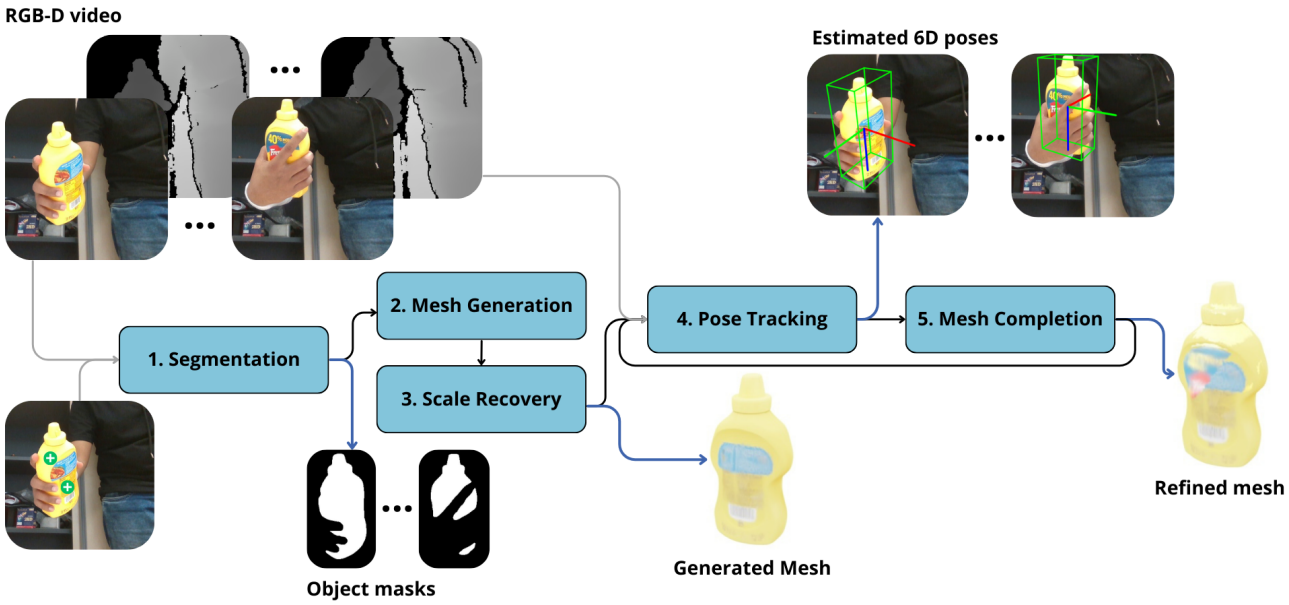

Figure 2: Overview of the proposed pipeline. Given an RGB-D video sequence and a segmentation prompt (two points on the object), the method proceeds through five stages: (1) object segmentation, estimating masks on all frames; (2) single-view mesh generation from the first frame; (3) scale recovery to align the mesh to the correct physical size; (4) 6D pose tracking throughout the video; and (5) mesh completion using RGB-D observations, while tracking the object.

A different approach, known as mesh completion, uses image-to-3D generative models to produce a complete initial mesh from a single image, enabling immediate tracking from the very first frame and refining the mesh online as new observations arrive. UA-Pose [3] follows this paradigm, representing object geometry as a neural SDF trained from synthetic renders of the generated mesh and introducing an uncertainty-aware layer that explicitly distinguishes observed from generated regions, progressively replacing synthetic geometry with real observations during tracking.

3.4. Image-to-3D Generation

SAM-3D [1] is a single-image 3D foundation model that predicts object shape and texture with strong robustness to partial occlusion. Like all generative approaches, however, it produces scale-agnostic outputs and must infer geometry in unseen regions. Any6D [2] builds on this kind of output and addresses the scale ambiguity through a coarse-to-fine alignment between the generated mesh and the observed depth map, recovering metric scale before using the mesh for pose estimation.

4. Proposed Method

The proposed method is designed around two guiding objectives: minimizing the required input, while achieving near real-time tracking performance. The first leads directly to the formulation in Section 2, which reduces the problem to its essential requirements — an RGB-D video and a segmentation prompt on the first frame. The second drives our mesh refinement strategy to be geometric rather than learned. The pipeline is model-free and zero-shot, chaining five stages illustrated in Figure 2. For segmentation, mesh generation, and pose estimation, we exploit existing models adapted to our setting. Our main contributions lie in the scale recovery and mesh completion stages.

Object segmentation estimates the mask on the prompted first frame and propagates it across all frames using an off-the-shelf video segmentation method. Mesh generation relies on SAM-3D [1], whose robustness to partial occlusion is essential, as the first frame may not show the object from an ideal viewpoint. SAM-3D also provides a scale estimate, but it is too coarse for accurate pose estimation. Our scale recovery stage addresses this by searching for the correct scaling factor with two progressive rounds.

In each round, a coarse pose is first estimated with FoundationPose [4], a set of scale hypotheses is applied to generate candidate poses, each candidate is scored by FoundationPose, and the highest-scoring scale is retained. With the correctly scaled mesh, FoundationPose initializes the 6D pose on the first frame and tracks it throughout the video by refining the previous pose at each new frame.

The mesh completion stage refines the geometry and appearance of the mesh over time using the RGB-D observations accumulated during tracking. Each vertex carries a per-vertex confidence scalar $\hat{c}$, initialized to zero, that quantifies how much consistent evidence has been gathered for its position. Every $n_{\text{comp}} = 10$ frames, the current depth map is preprocessed and unprojected into the mesh coordinate system under the estimated pose, yielding observed 3D point and color estimates.

Before any update is applied, the quality of the pose estimate is assessed. If the overlap between the projected mesh silhouette and the observed segmentation mask is too low or the median distance between observed points and the mesh is too large, the frame is discarded or confidence values are penalized.

Each observed point is assigned a confidence proportional to its distance from the mask boundary: points too close to the boundary, where mask inaccuracies are most severe, are discarded, while interior points receive higher confidence. Each point is then matched to a nearby mesh vertex. If the observation is geometrically consistent with the stored vertex position the vertex position, color, and confidence are updated by confidence-weighted averaging, and $\hat{c}$ is accumulated. Inconsistent observations instead decrease $\hat{c}$. A vertex is committed to the mesh geometry only once $\hat{c}$ reaches the commit threshold, ensuring that only positions supported by consistent evidence across multiple frames are incorporated, as illustrated in Figure 3. Vertices that never accumulate sufficient evidence retain their generative prior positions, preserving the complete initial shape on unobserved parts.

Because point assignment favors nearby vertices, some vertices are rarely reached through normal updates. Among these, spurious protrusions are identified as vertices that project inside the mask but lie closer to the camera than the observed depth — meaning they should be visible, but are not. These vertices are progressively penalized and eventually relocated to the observed surface. Finally, at the end of each update, smoothing is applied to reduce depth noise, and FoundationPose is reinitialized with the refined mesh.

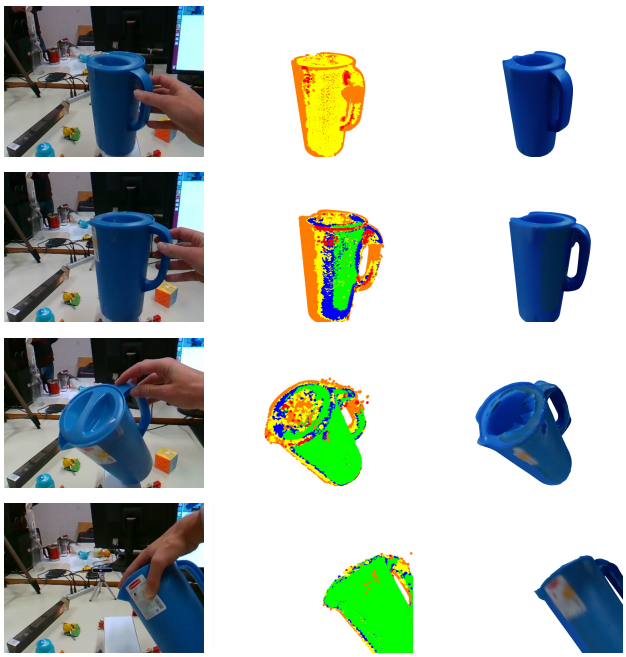

Figure 3: Evolution across frames. Each column shows the observed RGB image, the confidence-weighted point cloud, and the rendered mesh. Colors encode confidence from lowest to highest: red, orange, yellow, blue, green.

5. Experiments and Results

5.1. Dataset

We evaluate on HO3D v3 test set, an RGB-D dataset of hand-object interaction sequences featuring four YCB objects across 13 videos averaging approximately 1500 frames, which has become the standard benchmark for zero-shot model-free methods, having been adopted by the competing approaches [2, 3]. The dataset is well suited to our setting for three reasons: objects are frequently occluded by the manipulating hand or move partially out of frame, challenging both pose estimation and mask propagation; objects are actively rotated throughout each sequence, progressively revealing new surface regions that are informative for mesh completion; and the first frame often provides lim-

Table 1: Average results across all videos for pose and geometry metrics.

Method	ADD ↑ (%)	ADD-S ↑ (%)	CD ↓ (cm)	3D IoU ↑ (%)
UA-Pose	37.930	68.133	0.832	53.628
Any6D	38.158	70.638	0.559	78.793
Ours	56.397	80.273	0.400	**86.230**
OursComp	**60.516**	**82.031**	**0.352**	**86.230**
Baseline	68.976	84.490	0.225	91.587

ited, partially occluded object visibility, making initialization particularly challenging for zero-shot methods.

5.2. Metrics

Pose accuracy is evaluated through ADD and ADD-S, while mesh quality is analyzed with Chamfer Distance, 3D IoU, and PSNR. ADD measures the mean point-to-point distance between the ground-truth mesh transformed under the estimated and ground-truth poses, isolating pose error independently of mesh quality, while ADD-S replaces exact correspondence with nearest-neighbor matching, making it invariant to symmetric objects. Since the objects in HO3D are symmetric or near-symmetric, both metrics are applied to all objects and are reported as the Area Under the Curve (AUC) of the cumulative success rate over thresholds swept up to ten percent of the object diameter, following standard practice [2, 3].
Chamfer Distance and 3D IoU are evaluated at the first frame, where the pose estimate is most reliable; for methods with completion enabled, Chamfer Distance uses the final refined mesh re-evaluated under the first-frame pose. Chamfer Distance measures geometric accuracy through a bidirectional formulation that penalizes both missing and spurious surface regions, while 3D IoU assesses scale recovery independently of point-level geometry by comparing the oriented bounding boxes of the two transformed point clouds. PSNR complements these geometry metrics by evaluating texture fidelity through a pixel-level comparison of a mesh rendering against the reference image, and is computed throughout the full sequence using the current mesh.

5.3. Results

Five configurations are evaluated in Table 1, all sharing the same segmentation masks and FoundationPose estimator, so differences in performance directly reflect the contribution of mesh quality. The **Baseline** runs FoundationPose on the ground-truth CAD model and represents an upper bound where object geometry is fully known. **Any6D** [2] and **UA-Pose** [3] are model-free but not fully zero-shot: they eliminate the need for a CAD model but still require an external reference image at initialization. **Ours** and **OursComp** are the proposed fully zero-shot configurations, differing only in whether the mesh completion stage is enabled. All reported timings were measured on an NVIDIA RTX A5000 GPU.

Pose Accuracy and Mesh Geometry. OursComp achieves the best ADD and ADD-S AUC, closing respectively within 8.5 and 2.5 percentage points of the model-based Baseline despite operating with no prior knowledge of the object. Even without completion, Ours surpasses both Any6D and UA-Pose by a substantial margin. The higher-fidelity geometry produced by SAM-3D [1] is the principal driver of this performance advantage. Chamfer Distance confirms that geometry quality and pose accuracy move in the same direction across all configurations, with our methods consistently achieving lower reconstruction error. Pose tracking runs FoundationPose for all methods and takes approximately 0.03 s per frame (33 fps).

Scale Recovery. 3D IoU, computed on the initial mesh before any completion, isolates scale recovery quality independently of the completion stage. Ours outperforms both Any6D and UA-Pose and closely approaches the Baseline, with the gap to UA-Pose being especially

pronounced. Both Any6D and UA-Pose rely on the same mesh generation model for their initial reconstruction, which produces lower-quality geometry than SAM-3D and introduces proportional scale distortions. Any6D compensates through an elaborate per-axis scale recovery strategy that, despite starting from lower-quality geometry, achieves a good initialization, though at a higher computational cost (approximately 42 s, compared to 20 s for our method). UA-Pose instead adopts a simpler scaling approach (approximately 28 s), but the resulting initialization is of lower quality and constitutes its primary bottleneck, degrading all downstream metrics. These are one-time costs that do not affect per-frame tracking and add to the mesh generation step, which takes approximately 40 s for both of those methods.

Effect of Completion. Enabling the completion stage yields consistent gains across all metrics. The improvement is most pronounced on sequences where the initial SAM-3D mesh is weakest. UA-Pose also applies online completion via neural SDF retraining, yet reaches roughly the same pose accuracy as Any6D. This shows that completion has a measurable effect, but that a poor scale initialization degrades the starting mesh to a point where completion alone cannot fully compensate.

Appearance and Efficiency. Table 2 reports mean PSNR over all sequences, where both *Initial* and *Current* columns are included to isolate the contribution of completion to appearance quality. Completion improves appearance for both methods that apply it. Notably, OursComp surpasses the ground-truth Baseline in PSNR because the geometric vertex update adapts texture directly from video frames under actual scene lighting, rather than relying on the fixed texture of a CAD model. The two completion strategies differ markedly in computational cost: despite OursComp triggering far more updates on average (150 vs. 19), each update takes approximately 1.4 s, for a total of approximately 3.5 min, compared to approximately 236 s per update and 74 min for UA-Pose. Neither approach achieves real-time completion, but the geometric vertex update is substantially closer to practical deployment.

Table 2: Mean PSNR (dB) over all sequences. *Initial* uses the mesh before completion, *Current* uses the completion-updated mesh. The baseline renders the GT mesh at the GT pose.

Method	Initial ↑ (dB)	Current ↑ (dB)
UA-Pose	11.5	15.4
OursComp	14.7	**18.6**
Baseline	**15.3**	15.3

6. Conclusion

The results confirm that zero-shot model-free tracking is a practically viable alternative to model-based methods, with the proposed pipeline reaching near model-based accuracy with no prior knowledge of the object. The principal remaining bottleneck is the completion stage, where the geometric approach already demonstrates substantially lower overhead than neural alternatives, positioning it as the most promising direction for closing the gap to real-time operation. Achieving this would make the full pipeline deployable in interactive settings.

References

[1] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

[2] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6d: Model-free 6d pose estimation of novel objects. In *CVPR*, 2025.

[3] Ming-Feng Li, Xin Yang, Fu-En Wang, Hritam Basak, Yuyin Sun, Shreekant Gayaka, Min Sun, and Cheng-Hao Kuo. Ua-pose: Uncertainty-aware 6d object pose estimation and online object completion with partial references. In *CVPR*, 2025.

[4] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, June 2024.