IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

SilRef: Joint Visual Silhouette and Tactile Pose Optimization for Transparent Object Manipulation*

Jean-Baptiste Weibel[1,3], Clemence Dubois[2], Negar Layegh Khavidaki[1], Saifeddine Aloui[2], Mathieu Grossard[2], Markus Vincze[1] and Andreas Holzinger[3]

***Abstract*—Transparent objects are ubiquitous in laboratory automation settings, as liquids need to be visually controlled regularly. Automating laboratory processes would make the creation of small-batch medication feasible, thus making more personalized and better-targeted treatments more accessible. However, transparent objects present a major challenge for robust vision systems, in turn compromising their manipulation. Their appearance varies depending on the environment and depth sensors fail to capture their measurements. These objects therefore break central assumptions made by depth-based as well as render-and-compare pose refinement strategies. To ensure reliable pose estimation, we propose Silhouette-based object pose Refinement (SilRef), a novel pose refinement approach leveraging object silhouette detection and geometric cues, circumventing the need for depth maps or realistic rendering making it robust to environment change. Our proposed formulation directly optimizes the poses by gradient descent based on 3D models rendering and benefits from a large convergence basin. SilRef is evaluated on the Keypose dataset and the newly collected Tracebot In-Gripper dataset. Results show an improvement of 2.8x and 2.7x in Average Distance of Model Points-Symmetric (ADD-S@0.01m) when the object is standing on a surface and when the object is already grasped, respectively, compared to Megapose6D and ICP (Iterative Closest Point).**

***Index Terms*—Perception for Grasping and Manipulation; Deep Learning for Visual Perception; Deep Learning in Grasping and Manipulation**

I. Introduction

LABORATORY automation opens the door to affordable small batch manufacturing and therefore has the potential for significant improvement in health through more personalized treatment [1]. However, the automation in that field has been limited by the complexity of the manipulation of small to medium transparent containers.

Such containers are essential to enable constant supervision of the manipulated liquids, but they are a significant challenge for vision systems. In fact, transparent objects not only take on the appearance of their background, but are also very inaccurately estimated by depth-sensing methods [2]. This challenges standard manipulation pipelines as they either rely on direct grasp estimation [3], which relies on accurate depth, or first estimate the pose of the object to be grasped. For the latter approach, monocular object pose estimation has seen significant progress [4] but still needs to be complemented by an object pose refinement stage to achieve the best performance [5].

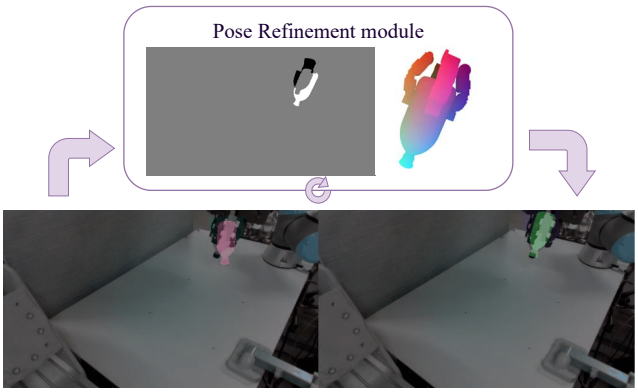

Fig. 1: Overview of SilRef. Object poses of transparent objects are iteratively refined based on their detection silhouettes and geometric cues (tactile sensor when the object is grasped or supporting surface when the object is standing).

Two strategies are employed for refining initial object pose hypotheses: depth-based refinement and render-and-compare. Iterative Closest Point (ICP) [6] is the most commonly used depth-based refinement method thanks to its simplicity and accuracy. ICP is however not applicable to transparent objects due to their depth-sensing challenge. Recently, render-and-compare methods have emerged as competitive alternatives [7] by using deep neural networks to score the current hypothesis and iteratively predict transformations to refine estimates. Accurately rendering transparent object models is, however, challenging, preventing meaningful hypothesis scoring and refinement in render-and-compare approaches.

As illustrated in Figure 1, we propose Silhouette-based object pose Refinement (SilRef), a method to circumvent those issues by optimizing the pose of objects based on geometric cues and detection silhouettes, which are comparatively more reliable for transparent objects. This approach only requires basic rendering to obtain the object silhouette. In a novel

Manuscript received: August 22, 2025; Revised: November 9, 2025; Accepted: December 16, 2025. This paper was recommended for publication by Editor Abhinav Valada upon evaluation of the Associate Editor and Reviewers' comments.

*The research leading to these results has received funding from EC Horizon 2020 for Research and Innovation under grant agreement No. 101017089, TraceBot. Corresponding author email address: jean-baptiste.weibel@boku.ac.at

[1]Vision for Robotics Laboratory, Automation and Control Institute (ACIN), TU Wien, Vienna, Austria

[2]Robotic Systems Architecture Laboratory, Universite Paris-Saclay, CEA, LIST, Palaiseau, France

[3]Human-Centered AI Lab, Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

formulation, SilRef combines the minimization of the distance between silhouette rays and the currently visible object points with geometric cues, which are either tactile sensing for in-gripper refinement, or supporting planes for standing objects. These geometric cues enable us to resolve ambiguities and fully constrain the pose optimization process. The poses are directly optimized using gradient descent. The process is illustrated in Figure 2. To support the evaluation of our approach, we also collected the novel Tracebot In-Gripper dataset consisting of 608 images of grasped transparent objects, as well as the corresponding output of tactile sensors. The dataset contains seven transparent objects under various poses and backgrounds (Fig. 3) and is used entirely for evaluation, as our method is training-free.

In summary, our contributions are:

- SilRef, a transparent object pose refinement method based on a novel silhouette- and geometry-based pose optimization formulation, that can flexibly include tactile sensing.
- the novel Tracebot In-Gripper dataset containing 608 images of grasped transparent objects with tactile sensing.
- a comparison to state-of-the-art object pose refinement methods applied to transparent objects to validate the approach.

In the following, we present the relevant state of the art in Section II, then present our proposed method in Section III. In Section IV, we present our experimental results before concluding in Section V.

II. Related Works

We present relevant works on transparent object pose estimation, object pose refinement and tactile-visual pose estimation.

A. Transparent Object Pose Estimation

Transparent object pose estimation has received renewed interest with the progress achieved by Deep Learning approaches. KeyPose [8] proposed to use a stereo camera and detect keypoints in both images followed by a Procrustes analysis to recover the 6D pose of transparent objects. TGF-Net [9] proposed a monocular-only approach that explicitly leverages edges from the image to support the pose estimation. Finally, ReFlow6D [10] propose a physics-inspired approach instead, predicting a refractive flow that models the change of direction of the light ray by a given object, and use it as an intermediate representation to predict the final 6D object pose. Despite significant progress, monocular pose estimation of transparent objects does not produce accurate enough pose estimation for manipulation, leading to the need for pose refinement methods

B. Object Pose Refinement

Object Pose Refinement aims to produce more accurate and reliable pose estimates from initial object pose hypotheses. Iterative Closest Point (ICP) [6] has long been used for that purpose, especially since the advent of affordable off-the-shelf 3D sensors. VeRefine [11] proposed to perform physics-informed pose refinement on top of ICP, sampling through the space of pose hypothesis using a Monte-Carlo Tree Search and alternating between steps of depth and normal alignments with step of physical simulation for each hypothesis to produce physically sound and accurate estimates. In [12], a reinforcement learning scheme is used to train an agent that iteratively refine poses, and includes physics supervision through losses applied to the agent training. All these methods rely on depth sensing to refine the object pose, leading to suboptimal performance on transparent objects due to their depth sensing challenge. DeepIM [13] demonstrated that a deep neural network could iteratively predict pose corrections by taking as input a rendered color image of an object model together with the scene image. Megapose [7] later built on this approach and demonstrated that it was possible to train a model capable of handling previously unseen objects with a varied enough training set. Finally, FoundationPose [14] pushed that idea further combining it with a pose scoring network to propose a framework capable of pose estimation and pose tracking through refinement, thanks to a very large-scale training dataset to generalize. That category of methods requires rendering to compare the current pose estimate to the image, which is challenging for transparent objects as their appearance depends on their background.

C. Tactile and Tactile-Visual Pose Estimation

Tactile and Tactile-visual pose estimation has been explored before in robotics scenarios. Advanced tactile sensors are used to extract object shape information. [15] and [16] use such a sensor together with a tactile sensor simulator and extract shape silhouettes directly from the tactile sensor that can be matched to ones extracted by the simulator on the object model to predict the final pose. Alternatively, [17] proposes to combine tactile and depth information into a more complete point cloud that can be used for pose estimation. In [18], the authors propose a similar idea, combining depth and tactile sensors to form a more complete point cloud, and use it as input to a network fusing information from point clouds and RGB images to predict the final pose. Finally, [19] proposes to first learn to complete the object shape using RGB-D and tactile sensors before using the completed shape to predict the pose.

All these approaches rely either on depth data, limiting their reliability for transparent objects or on the availability of advanced tactile sensors that estimate shapes from contact, limiting their applicability.

III. SilRef: Visual-Tactile Transparent Object Pose Refinement

We introduce SilRef, the first method to refine transparent object poses. The method relies on a gradient-descent optimization combining visual and geometric information easy to obtain for transparent objects. In particular, a novel ray-point formulation is used to leverage detection silhouettes in the optimization. Furthermore, the pose optimization is fully constrained using one of two geometric cue: we use tactile sensing to refine the object's in-hand pose to support bi-manual regrasping and accurate object placement, and we

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

use the supporting surface information to enhance transparent object pose before it is picked up. All optimization losses are illustrated in Figure 2. In the rest of this section we describe both the general optimization scheme, and the visual and geometric supervision used in our optimization.

A. Direct Object Pose Optimization

We propose to frame the task of transparent object pose refinement as a direct gradient-based joint optimization of the pose of each object in the scene. The pose hypotheses given as input to the refinement system are expected to be close to the correct pose leading us to approximate the problem as a convex optimization. We create a virtual scene by placing the known 3D models of the set of objects $O = \{o\}$ under the initial pose hypotheses. As this transformation is differentiable, when placing constraints on the 3D models points, the gradient can flow all the way back to the pose estimate and update it accordingly. We simply represent the pose parameters as a quaternion, transformed in a rotation matrix as needed and a translation vector in the optimization process. We also complement our optimization scheme with a differentiable renderer. By rendering the virtual scene, we can relate each image pixel to the corresponding point on the 3D model surface, given the current pose estimates. This process naturally handles self-occlusion and occlusion between scene objects based on the depth of each 3D model point, as only the closest point to the camera frame will be considered for each pixel. This property lets us directly relate 3D model properties to 2D image properties and therefore compare it to estimation performed on the input RGB image. In particular, detection silhouettes are already available, as they are necessary for the initial pose estimation process and are easy to render efficiently, even for transparent objects. This circumvents the need for a complex rendering pipeline or accurate depth sensing to compare against. In case the optimization is not convex (poor pose hypotheses or incorrect detection silhouette), the detected and rendered silhouette will not align well, making it straightforward to detect such cases based on the final intersection over union of the two silhouettes. We detail how we supervise the pose optimization based on the detection silhouette in Section III-B. As constraints on the object 3D points can be directly related to the pose estimate, they can also create geometric constraints to ensure contact or avoid collision. We describe how we supervise the pose optimization based on geometric constraints in Section III-C. While more elaborate schemes could be investigated, we propose to simply weigh all losses considered equally after ensuring they are all within similar range.

B. Visual supervision

We propose to leverage object silhouettes for the pose optimization. Object detection and segmentation has indeed proven reliable even for transparent object instances despite their lack of features [20]. Given an object model and an initial 6D pose hypothesis, we optimize the pose by comparing the rendered silhouette to the detection silhouette estimated from the RGB image.

While a differentiable renderer could be used to directly compare the two silhouette images, we propose an alternative geometric formulation. We minimize the distance between the rays $\mathcal{R}_o = \{\vec{r_o}\}$ (normalized) extracted from the detection silhouette of the object o and the points $\mathcal{P}_o^* = \{\vec{p_o}\}$, the visible subset of the point $\mathcal{P}_o$ of the corresponding object model after rendering, computed for every object in the scene:

$$l_{silhouette} = \sum_O \left(\frac{1}{2|\mathcal{P}_o^*|} \sum_{\vec{p_o} \in \mathcal{P}_o} \min_{\vec{r_o} \in \mathcal{R}_o} \|p_{\|\vec{r},o} - \vec{p_o}\| + \frac{1}{2|\mathcal{R}_o|} \sum_{\vec{r_o} \in \mathcal{R}_o} \min_{\vec{p_o} \in \mathcal{P}_o^*} \|p_{\|\vec{r},o} - \vec{p_o}\| \right) \tag{1}$$

with $p_{\|\vec{r},o} = (\vec{r_o}.\vec{p_o})\vec{r_o}$ the projection of $\vec{p_o}$ on the $\vec{r_o}$. This ensures that every ray from the detection silhouette will be close to an object point, and every object point will be close to a ray, such that every ray and point are explained after optimization.

This formulation leverages all the silhouette pixels available without computing unnecessary distance between every pixel of the silhouette and projected image, in particular the distance between background pixels, as would be done by directly comparing silhouette images. It has the added benefit of still being valid even when no overlap happens between both silhouettes.

C. Geometric supervision

During the rendering process, a small variation of the object pose along the optical axis of the camera (its depth in the camera frame) will lead to little to no difference in the silhouette produced. Compounded by the potential noise in the detection silhouette, which itself may be a bit larger or smaller than the actual object, this issue leads to unreliable depth estimates for the object pose when using the visual supervision alone.

As such we propose to complement it using simple geometric constraints to obtain robust and accurate poses. We propose two different constraints that can be used and that together cover the majority of manipulation scenarios. On one hand, when the object is already in the gripper, we propose to complement visual supervision with tactile sensing. On the other hand, when the object is standing, we propose to use the information about its supporting surface. We describe both constraints in the next sections.

1) Tactile sensing: We supplement the visual supervision with tactile information when the object is in the gripper. As many tactile sensing technologies exist with varying output, we propose to supervise the optimization using information that all can provide to ensure generality of the method. The optimization is therefore based on contact points extracted from tactile sensors, whose 3D positions relative to the objects to optimize are inferred through the kinematics of the robot arm and gripper and relative pose to the camera. The object pose must therefore guarantee the existence of the object surface in the vicinity of the contact point sensed, that is the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

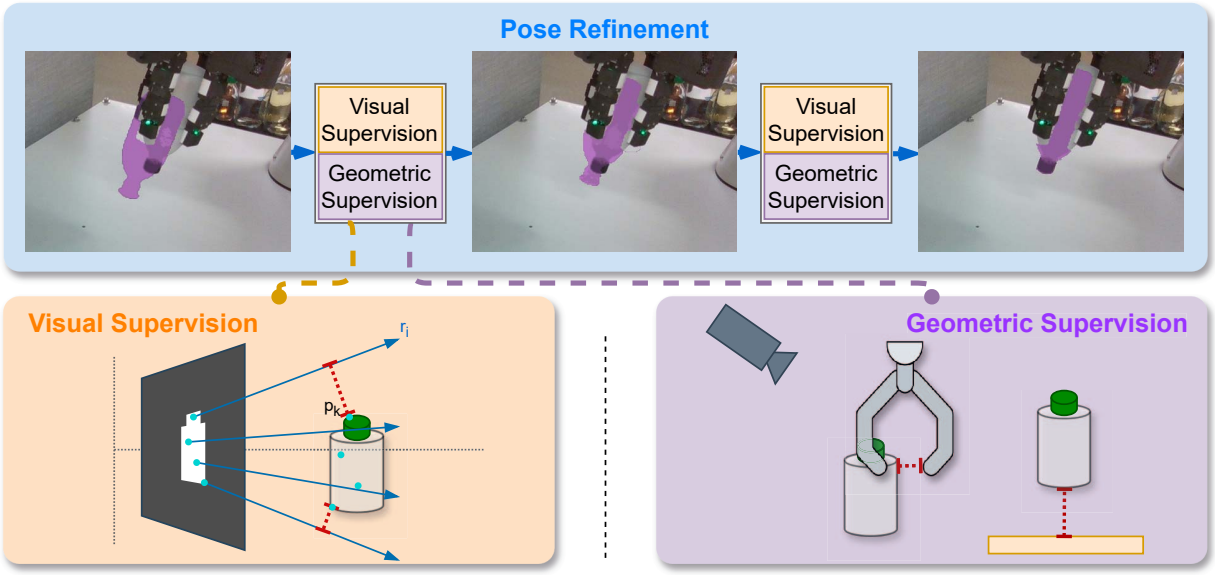

Fig. 2: Illustration of the optimization process. Top: the pose of the transparent object is iteratively refined based on visual and geometric supervision. Bottom left: the visual supervision (Eq.1) minimizes the distance between 3D model points and detection silhouette rays illustrated as red dotted lines. Bottom right: the geometric supervision minimizes the distance between the gripper and the 3D model (Eq.2) or the distance between the 3D model and the closest supporting surface (Eq.4)

position of the tactile sensor. Such a criterion can simply be implemented using an L2-loss, with P the set of points of the object model and T the set of contact points detected by the tactile sensor:

$$l_{tactile} = \sum_{t \in T} \min_{p \in P} \|p - t\| \tag{2}$$

With this loss, the distance between the different contact points detected by the tactile sensor and the object surface is minimized.

The position of contact points is however subject to noise, coming from the sensor itself and the kinematics (low sensor spatial resolution, imprecise gripper joint pose estimation, imprecise arm calibration). The configuration of the contact points might not constrain all degrees of freedom of the pose if they are all coplanar, as is the case with a parallel grasp. Coplanar and noisy contact points can lead to the L2-loss pushing away from the correct position along the unconstrained axis to ensure a surface is close to their noisy position. To address this issue and increase the robustness of the optimization, we propose to adapt the loss defined in Equation 2 based on the eigenvalues of the covariance matrix of the contact points positions. If the two largest eigenvalues cover more than 95% of the variance, that is to say the contact points are coplanar, we first project the points onto the contact points plane, and minimize the distance of the projection of the object points to the contact points. With $d_{plane} = (\vec{p} - \vec{t_{plane}}).\vec{n_{plane}}$ the distance of point p from the object model to the plane defined by the point and normal vector $(t_{plane}, \vec{n_{plane}})$, and $p_{plane} = p - d_{plane}\vec{n_{plane}}$ the projection of p onto that plane:

$$l_{tactile} = \sum_{t \in T} \min_{p_{plane} \in P} \|p_{plane} - t\| + \frac{d_{plane}}{10} \tag{3}$$

We still consider the distance of the object points to the plane, but down weigh that value by a factor of ten as contact points do not give much information along that axis in that case.

2) Supporting surface: Before the object has been grasped and is still standing in the scene, we propose to leverage information about its supporting surface. Defining the surface as a point and a normal, we enforce the object to stay above it but prevent the object from going through it. Due to the complex interactions that can happen between objects, for example if the object is attached to a stretched out tube, we do not enforce a strict contact with the supporting surface.

We formalize this constraint by projecting all points of the 3D model along the direction of the surface and only consider the closest point to the supporting surface in that projection, which we call the critical point. We enforce that this critical point remains close to the supporting surface (close to zero after the projection), and penalize it more strongly if that distance is negative (below the surface, meaning colliding) than if positive (hovering above). Formally, the loss is defined as, with $p_{crit} = \min_{p_{\perp surf} \in P}(p_{\perp surf})$ and $p_{\perp surf} = (\vec{p} - \vec{p_{surf}}).\vec{n_{surf}}$ the projection of the point $\vec{p}$ onto the surface defined by its normal $\vec{n_{surf}}$ and a point $\vec{p_{surf}}$:

$$l_{surf} = \begin{cases} Huber(p_{crit}, 0) \text{ if } p_{crit} < 0 \\ \frac{1}{\beta} Huber(p_{crit}, 0) \text{ if } p_{crit} > 0 \end{cases} \tag{4}$$

The pose is therefore less penalized by a factor β when the lowest projected point is above the surface than when it is below. It is of note that while the criterion models the

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

supporting surface as a plane (point and normal direction), the actual surface does not have to be.

While we strive for general applicability and minimal scene knowledge, SilRef is also flexible as further criterion can easily be included when available. For example, a loss can be added to enforce the alignment of the closest stable axis of the object axis with the surface normal, or even reduce the degrees of freedom of the pose to only allow poses on the supporting surface itself.

IV. Experiments

In this section, we evaluate the performance of SilRef (i) on grasped transparent objects scenarios in Section IV-B using the Tracebot in-gripper dataset we collected (ii) on standing transparent objects scenarios in Section IV-C using the established KeyPose dataset [8] (iii) with and without geometric supervision in Section IV-D1 (iv) under varying silhouette quality in Section IV-D2 to evaluate the robustness of our method. We compare our approach to commonly used state-of-the-art methods representative of the two main refinement strategies: Megapose6D [7] for render-and-compare and ICP [6] for depth-based refinement.

A. Experimental Setup

We now detail the parameters and methodology used to evaluate our proposed approach.

1) Optimization settings: SilRef is evaluated using the point-ray loss defined in Equation 1 combined with either the tactile contact points loss defined in Equation 2 for the Tracebot in-gripper dataset or the supporting surface of Equation 4. All losses are weighed equally. The supporting surface loss is computed in centimeters to obtain values in a similar range. We use the Adam optimizer with a learning rate of 0.005 and set the number of iterations to a 100 to ensure the method has reached convergence. Speed was not the focus of this work and implementation, convergence criterion and early stopping could therefore be optimized. The number of iterations per second has a high variance as it depends on the rendering resolution, the number of objects (every objects in the scene are jointly optimized) and the number of rays in each detection silhouette (distance of the object to the camera) Our unoptimized approach performs 15 refinement iterations per second on average on an Nvidia 4070 Ti Super (66ms per iteration). In practical scenarios, we observed that the method required less than 30 iterations to sufficiently converge to grasp objects. The runtime was comparable or faster than Megapose6D.

2) Evaluation Methodology: To evaluate the performance of object pose refiners, we sample rotation and translation noise to add to the groundtruth pose. We first sample a random rotation axis and rotate the object by $5\,\mathrm{deg}$, $20\,\mathrm{deg}$ or $40\,\mathrm{deg}$. As a second step, we translate the object along a random direction by 25%, 50% or 75% of the object size. The final noise sampling consists of every combination of rotation and translation, leading to nine estimations per object pose. The same rotation and translation noise is used over each experiment and method.

Fig. 3: Tracebot in-gripper dataset objects and example scenes.

3) Datasets: Two datasets are used for evaluation. We collect the Tracebot In-Gripper, the first transparent object dataset that includes pose annotation and tactile sensing data to evaluate in-gripper pose refinement. We also use the Keypose dataset as it is a benchmark for transparent object pose estimation.

The Tracebot In-Gripper dataset, our newly introduced dataset, is used to demonstrate the usefulness of integrating tactile sensing in our transparent object pose refinement. It consists of 19 scenes, each with two static cameras capturing 16 images of a robot arm holding a transparent object in a different pose each time for a total of 608 images. The groundtruth is obtained using the 3D-DAT [21] annotation toolkit, which let users leverages camera poses and multi-views to refine object poses based on their reprojection in the image, making it suitable for transparent objects. Besides the images, we also collect the positions of contact points occurring between the transparent object and a four-finger gripper equipped with 13 tactile sensors (one sensor per phalanx, three phalanges per finger, and one sensor for the palm) [22]. A simple processing is used to extract a single contact point per sensor. Both a uniform and a textured background are used to further challenge object detection and segmentation methods, and the objects chosen cover different sizes and different fill levels to vary the type of transparency. It should also be noted that significant occlusion of the object of interest occurs in these scenes because of the fingers of the gripper. The objects used in that dataset as well as some example images are illustrated in Figure 3.

We also test our approach on the test set of the KeyPose dataset [8] as it is already an established benchmark for transparent object pose estimation. The test set consists in 56 scenes of 80 images with a single object, collected with a camera attached to a moving robot arm. 14 different objects are captured in 4 different poses each. The scenes are illustrated in Figure 4.

4) Silhouettes estimation: For all dataset, unless specified otherwise, silhouettes are obtained using SegmentAnything2 [23] (SAM2). For the Tracebot In-Gripper dataset, point prompts are manually created on the object in the first image of the sequence. For the Keypose dataset, point prompts are created by sampling three points inside the eroded ground truth

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Fig. 4: Example images from the KeyPose dataset [8]

			ADD-S Threshold (%)		
Method	Input	Dataset	0.01	0.02	0.05
Megapose6D [7]	RGB	TBot	9.1	23.8	52.8
ICP [6]	D	TBot	19.9	47.0	74.7
Ours (Tactile)	RGB	TBot	**54.0**	**82.7**	**92.3**
Megapose6D [7]	RGB	Keypose	18.4	30.5	48.8
Ours (Surface)	RGB	Keypose	**51.3**	**87.4**	**99.2**

TABLE I: Main results on the Tracebot In-Gripper and Keypose dataset.

masks (to avoid prompts at the border). The annotations of every other image in the sequence are obtained by letting SAM2 automatically propagate the first image annotation through cross-attention blocks between its memory bank and the encoded image tokens. This means the vast majority of silhouettes are obtained without direct prompting of the corresponding RGB images.

5) Metrics: To evaluate the final pose quality, we report the Average Distance of Model Points-Symmetric (ADD-S) [24] score for the pose estimation, that is the distance between points $\mathcal{P}$ of the object model after being transformed by the pose P^{pred} and the nearest point of $\mathcal{P}$ transformed by the groundtruth pose P^{GT}:

$$e_{ADD} = \text{mean}_{p \in \mathcal{P}} \| P^{pred} p - \mathcal{NN}(P^{GT} \vec{p} \in \mathcal{P}) \| \quad (5)$$

We chose this metric over ADD [24] as it better represents the pose quality when considering symmetric objects. We also report the percentage of pose with an ADD-S below various thresholds to better represent the behavior of pose estimators. We pick fixed thresholds in centimeters rather than percentage of object sizes, as we consider it more relevant to manipulation.

B. In-Gripper refinement results

We first evaluate SilRef when refining the pose of objects already inside grippers, to enable better support for accurate object placement or bimanual operations. We compare our proposed approach leveraging silhouettes and contact points from a tactile sensor to state-of-the-art methods, including depth-based approach ICP [6] and render-and-compare approach Megapose6D [7]. The percentage of estimates with an ADD-S below a set of thresholds is presented in Figure 5 and in Table I.

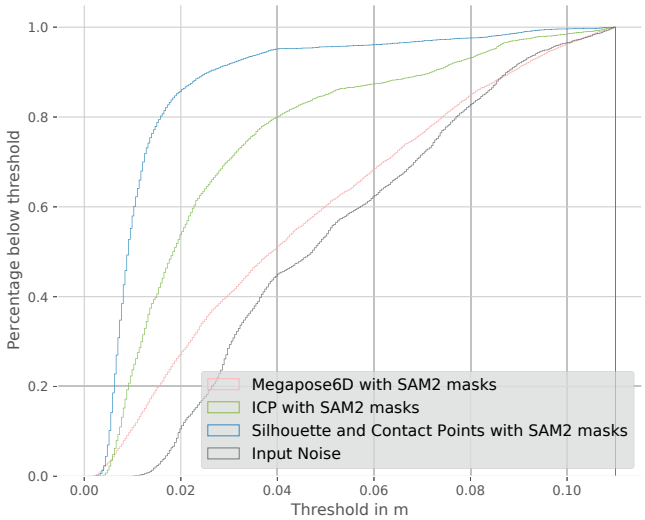

Fig. 5: Comparison of ADD-S threshold curves of object pose refinement methods on the Tracebot In-Gripper dataset.

SilRef clearly outperforms ICP, due to the inaccurate depth of transparent objects, and Megapose6D, whose performance is hindered by inaccurate rendering of transparent objects. As shown in Table I, our method reaches an ADD-S value of less than 1cm in the majority of the experiments (54%), which is accurate enough to support a wide range of manipulation.

C. Supporting surface refinement results

We then evaluate our proposed approach in cases where objects are lying or standing on a surface, to enable picking up such objects. We compare our proposed approach leveraging silhouettes and our supporting surface loss to the state-of-the-art method Megapose6D [7]. ICP is unfortunately not applicable on this dataset as no depth information is provided. The percentage of estimates with an ADD-S below a set of thresholds is presented in Figure 6 and in Table I.

SilRef improves by a factor of 2.8 in terms of ADD-S@0.01m over Megapose6D [7], while keeping a similar runtime.

D. Ablation study

We now present experiments testing the usefulness of our proposed geometric cues, the robustness of our approach to varying silhouette estimation methods and the impact of the scene background. All results are compiled in Table II.

1) Impact of geometric supervision: We evaluate the usefulness of the geometric cues chosen in our approach. Figure 7 compares the ADD-S threshold curves using the contact points sensing alone, the silhouettes alone and both combined.

We can see that only the combination of both achieves an ADD-S value below 1cm for a significant portion of the experiment, which is necessary for the manipulation of transparent objects. Using only contact points reach a slightly higher proportion of prediction with an ADD-S below 5cm.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

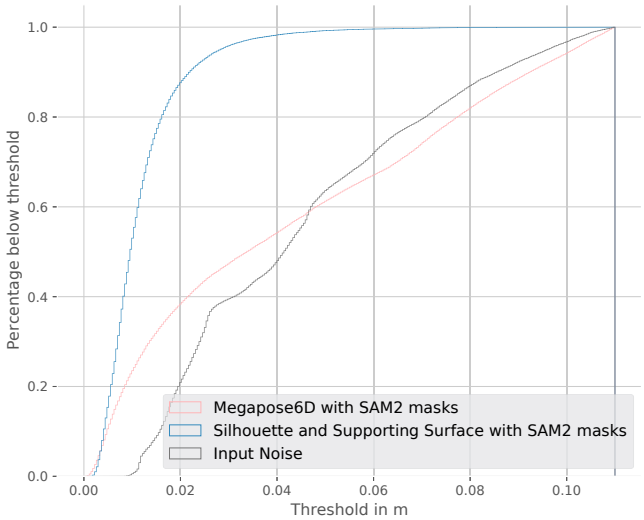

Fig. 6: Comparison of ADD-S threshold curves of object pose refinement methods on the KeyPose dataset.

			ADD-S Threshold (%)		
Silhouette	Geom. cue	Dataset	0.01	0.02	0.05
-	Tactile	TBot	32.5	78.1	**99.4**
Yolov8	-	TBot	2.3	15.7	50.8
Yolov8*	-	TBot	11.6	48.9	79.7
GT	-	TBot	29.2	62.9	89.5
SAM2	-	TBot	15.1	55.4	85.3
SAM2	Tactile	TBot	**54.0**	**82.7**	92.3
SAM2	-	Keypose	29.3	63.5	93.8
SAM2	Surface	Keypose	**51.3**	**87.4**	**99.2**

TABLE II: Ablation Studies

This is due to the strong geometric supervision provided by contact points, which prevents objects from drifting away but does not provide enough information to fully constrain the pose, preventing a higher proportion of prediction with an ADD-S value below 1cm.

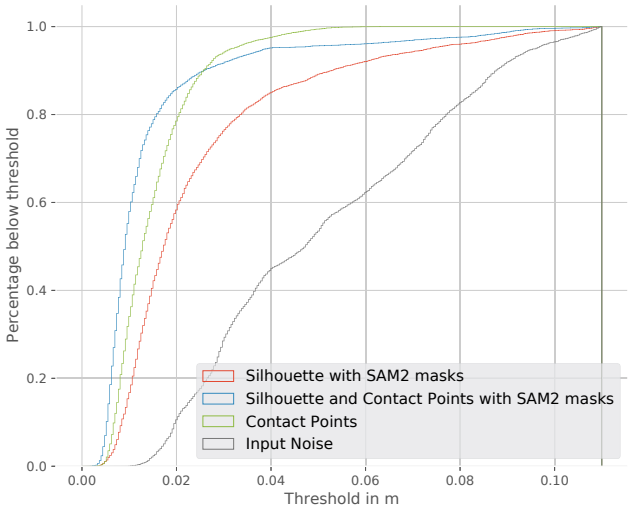

Fig. 7: Comparison of ADD-S threshold curves depending on the in-hand geometric cue on the Tracebot In-Gripper dataset.

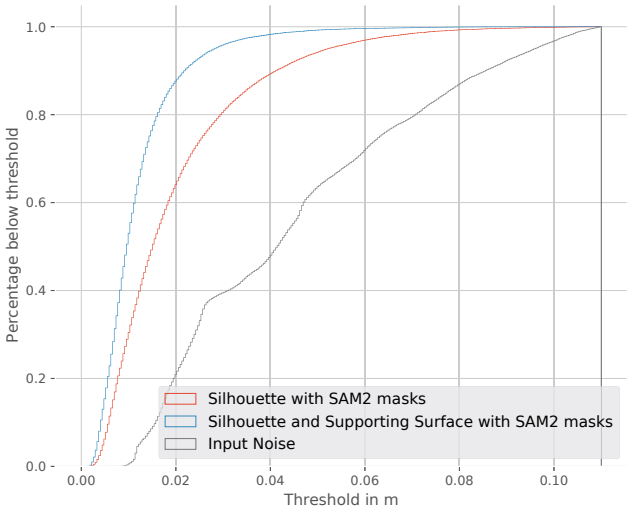

Fig. 8: Comparison of ADD-S threshold curves depending on the surface geometric cue on the KeyPose dataset.

Figure 8 compares the ADD-S threshold curves when using the silhouettes alone and when combined with the supporting surface. It demonstrates the usefulness of the supporting surface cue.

2) Impact of silhouette quality: Figure 9 presents the impact of the silhouette estimator on our method on the Tracebot In-Gripper dataset. Results with the ground truth silhouettes are compared to the results obtained with SAM2 silhouettes and with Yolov8 [25] silhouettes. Yolov8 was chosen to evaluate the impact of the silhouette's quality on our approach as it is a popular model which produces masks of significantly lower quality than SAM2. As Yolov8 combines detection and segmentation, we report the results obtained on images with a correct detection (indicated by a *) and the results on all images (for which undetected objects use the ADD-S value of the input noise).

SAM2 provides the highest quality silhouettes, despite not being prompted on the test image but only the first image of each scene. Yolov8 on the other hand, predicts silhouettes at a low resolution, leading to less accurate boundaries. Moreover, occlusions by the gripper's finger are not always accurately predicted by Yolov8. Our method still presents a robustness to these silhouette artifacts as shown by the results when only considering correct detections in Table II. Accurate boundaries remain the most important factor. Indeed, after projection, a dilated silhouette is equivalent to a silhouette of the same object closer to the camera, and vice-versa for an eroded silhouette.

Looking at the impact of the background on our results, we also notice that segmentation methods are quite robust to more complex and textured backgrounds in that scenario. Indeed, the average ADD-S score goes from 0.031m to 0.028m when comparing uniform backgrounds and textured backgrounds using SAM2 masks.

V. Conclusion

We presented SilRef, an approach that performs transparent object pose refinement by directly optimizing the pose based

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

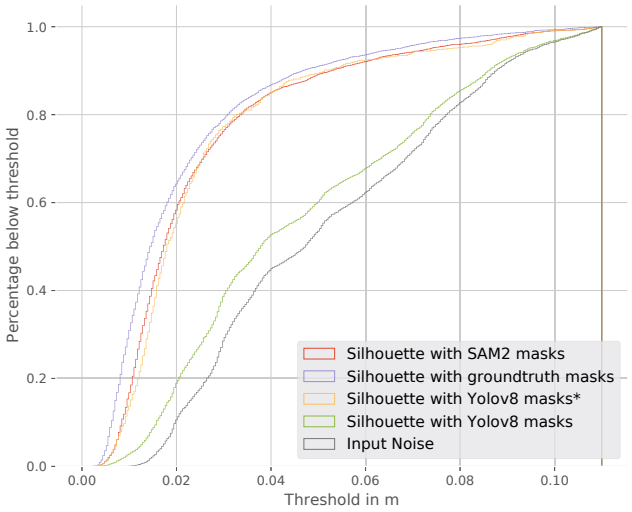

Fig. 9: Comparison of ADD-S threshold curves for different silhouette prediction methods.

on the detection silhouette combined with a geometric cue, either contact points provided by tactile sensors or information about the supporting surface. We demonstrated that such an approach outperforms the two categories of general object pose refinement methods, that is depth-based approaches and render-and-compare approaches. In particular, our proposed approach outperforms those methods by a factor of 2.7 in ADD-S@0.01m on the Tracebot In-Gripper dataset that we collected and by a factor of 2.8 on the Keypose dataset. More than half of the objects evaluated reach an ADD-S value below 0.01m, demonstrating that our work opens the door to more robust, real-world robotic manipulation of transparent objects in safety-critical domains such as laboratory automation.

Future work will investigate how to adapt such methods in the presence of uncertainty, including silhouette confidence and unknown objects in the scene. Furthermore, the presented approach will be extended to better address opaque objects and take advantage of texture information, leveraging edges to work toward a general solution for object pose refinement.

VI. Data Availability Statement

The dataset is available at https://doi.org/10.48436/3xejy-cws13

References

[1] C. Naugler and D. L. Church, "Automation and artificial intelligence in the clinical laboratory," *Critical Reviews in Clinical Laboratory Sciences*, vol. 56, no. 2, pp. 98–110, 2019, pMID: 30922144. [Online]. Available: https://doi.org/10.1080/10408363.2018.1561640

[2] J.-B. Weibel, P. Sebeto, S. Thalhammer, and M. Vincze, "Challenges of Depth Estimation for Transparent Objects," in *Advances in Visual Computing*, G. Bebis, G. Ghiasi, Y. Fang, A. Sharf, Y. Dong, C. Weaver, Z. Leo, J. J. LaViola Jr., and L. Kohli, Eds. Cham: Springer Nature Switzerland, 2023, pp. 277–288.

[3] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, May 2021, pp. 13 438–13 444.

[4] S. Thalhammer, D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, and M. Vincze, "Challenges for Monocular 6-D Object Pose Estimation in Robotics," *IEEE Transactions on Robotics*, vol. 40, pp. 4065–4084, 2024.

[5] T. Hodaň, M. Sundermeyer, Y. Labbé, 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," *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[6] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. SPIE, Apr. 1992, pp. 586–606.

[7] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare," in *6th Annual Conference on Robot Learning*, Aug. 2022.

[8] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, "KeyPose: Multi-View 3D Labeling and Keypoint Estimation for Transparent Objects," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020, pp. 11 599–11 607.

[9] H. Yu, S. Li, H. Liu, C. Xia, W. Ding, and B. Liang, "TGF-Net: Sim2Real Transparent Object 6D Pose Estimation Based on Geometric Fusion," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3868–3875, June 2023.

[10] H. Gupta, S. Thalhammer, J.-B. Weibel, A. Haberl, and M. Vincze, "Re-Flow6D: Refraction-Guided Transparent Object 6D Pose Estimation via Intermediate Representation Learning," *IEEE Robotics and Automation Letters*, vol. 9, no. 11, pp. 9438–9445, Nov. 2024.

[11] D. Bauer, T. Patten, and M. Vincze, "VeREFINE: Integrating Object Pose Verification With Physics-Guided Iterative Refinement," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4289–4296, July 2020.

[12] ——, "SporeAgent: Reinforced Scene-Level Plausibility for Object Pose Refinement," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 654–662.

[13] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "DeepIM: Deep Iterative Matching for 6D Pose Estimation," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 657–678, Mar. 2020.

[14] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects," Mar. 2024.

[15] M. B. Villalonga, A. Rodriguez, B. Lim, E. Valls, and T. Sechopoulos, "Tactile Object Pose Estimation from the First Touch with Geometric Contact Rendering," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 1015–1029.

[16] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2Pose: Tactile object pose estimation from the first touch," *The International Journal of Robotics Research*, vol. 42, no. 13, pp. 1185–1209, Nov. 2023.

[17] J. Bimbo, S. Rodríguez-Jimenez, H. Liu, X. Song, N. Burrus, L. D. Senerivatne, M. Abderrahim, and K. Althoefer, "Object pose estimation and tracking by fusing visual and tactile information," in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, Sept. 2012, pp. 65–70.

[18] S. Dikhale, K. Patel, D. Dhingra, I. Naramura, A. Hayashi, S. Iba, and N. Jamali, "VisuoTactile 6D Pose Estimation of an In-Hand Object Using Vision and Tactile Sensor Data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2148–2155, Apr. 2022.

[19] H. Li, S. Dikhale, S. Iba, and N. Jamali, "ViHOPE: Visuotactile In-Hand Object 6D Pose Estimation With Shape Completion," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 6963–6970, Nov. 2023.

[20] J. Jiang, G. Cao, J. Deng, T.-T. Do, and S. Luo, "Robotic Perception of Transparent Objects: A Review," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2547–2567, June 2024.

[21] M. Suchi, B. Neuberger, A. Salykov, J.-B. Weibel, T. Patten, and M. Vincze, "3D-DAT: 3D-Dataset Annotation Toolkit for Robotic Vision," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 9162–9168.

[22] T. Ayral, S. Aloui, and M. Grossard, "Spectro-Temporal Recurrent Neural Network for Robotic Slip Detection with Piezoelectric Tactile Sensor," in *2023 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*, June 2023, pp. 573–578.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "SAM 2: Segment Anything in Images and Videos," https://arxiv.org/abs/2408.00714v1, Aug. 2024.

[24] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562.

[25] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics