Insert-One: One-Shot Robust Visual-Force Servoing for Novel Object Insertion with 6-DoF Tracking

Haonan Chang[1,2], Abdeslam Boularias[2], Siddarth Jain[1]

***Abstract*— Recent advancements in autonomous robotic assembly have shown promising results, especially in addressing the precision insertion challenge. However, achieving adaptability across diverse object categories and tasks often necessitates a learning phase that requires costly real-world data collection. Moreover, previous research often assumes either the rigid attachment of the inserted object to the robot's end-effector or relies on precise calibration within structured environments. We propose a one-shot method for high-precision contact-rich manipulation assembly tasks, enabling a robot to perform insertions of new objects from randomly presented orientations using just a single demonstration image. Our method incorporates a hybrid framework that blends 6-DoF visual tracking-based iterative control and impedance control, facilitating high-precision tasks with real-time visual feedback. Importantly, our approach requires no pre-training and demonstrates resilience against uncertainties arising from camera pose calibration errors and disturbances in the object in-hand pose. We validate the effectiveness of the proposed framework through extensive experiments in real-world scenarios, encompassing various high-precision assembly tasks.**

I. INTRODUCTION

For decades, researchers have been captivated by the pursuit of autonomous robotic assembly, with a particular focus on the insertion problem [1]–[10]. Commonly known as the Peg-in-Hole Insertion (PIH) problem, it involves assembling an insertion object into a stationary receptacle. Despite its prevalence in industrial settings, achieving complex, high-precision assembly in unstructured environments remains a formidable challenge [11]. Uncertainties stemming from variations in grasp alignment, object positions, discrepancies in parts, and calibration errors can result in failures and collisions with surfaces. In unstructured scenarios where both grasping and insertion are required for assembly tasks, particularly those with minimal tolerances, relying solely on meticulous calibration often proves insufficient. Moreover, these uncertainties evolve during the physical interactions between the robot and the object. Incorporating feedback systems can help mitigate uncertainties by providing real-time information during physical interactions.

One approach to tackle the insertion problem involves force sensing to determine the exact position of the receptacle. Assuming both the insertion object and the receptacle reside on a common plane, the object is navigated across the plane of the receptacle in a search pattern to maintain contact throughout the search. Measurements obtained from a force sensor allow for the deduction of the hole's position using contact space modeling [12], [13]. Nonetheless, methodologies that solely rely on force feedback require a successful search pattern [14] and are limited to scenarios where the alignment of the insertion object and the receptacle is initially approximate. Many approaches assume a rigid connection between the object being grasped and the robot [12], [15], ignoring the complexities of the grasping process. This oversight can affect the accuracy and effectiveness in unstructured settings, as it does not account for potential flexibility or movement between the object and the robot.

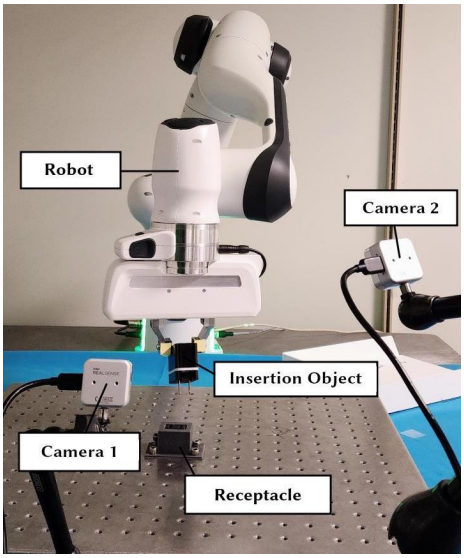

Fig. 1: Robotic setup for the insertion tasks. One object is designated as a stationary receptacle while the other is an insertion object placed randomly on a flat surface at the start. Initially, the robot must visually perceive and securely grasp the insertion object with its end-effector and then accurately position it so that it snugly fits into the receptacle.

One strategy to overcome the initial alignment challenge is to utilize visual feedback [16]–[20]. While this approach has demonstrated potential, many existing techniques rely on conventional visual servoing with manually designed features, which can result in instability. Recently, efforts have been made to incorporate reinforcement learning [15], [20]–[22] into insertion tasks, achieving better robustness. However, these learning approaches can pose risks to the robot and its equipment, especially in contact-intensive tasks. Additionally, they often require complex learning setups, detailed reward engineering, a large number of demonstrations, and may struggle with generalizing to new objects, tasks, and unstructured environments.

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. sjain@merl.com

[2]Department of Computer Science, Rutgers University, New Jersey, USA. This research was completed during H. Chang's internship at MERL.

In this paper, we present **Insert-One**, addressing the challenges of performing insertion for novel objects in a one-shot context, while accounting for errors stemming from grasping and visual pose estimation. Our approach integrates zero-shot model-based 6-DoF tracking [23], [24] into a two-stage vision force servoing pipeline, enabling seamless adaptation to new object categories. The key novelty behind our method lies in the use of only one RGB-D image as a demonstration of the task for visual alignment. Different from previous work [21], our demonstration is a single static image. During the visual alignment phase, we deploy tracking-based feedback control to synchronize the object's pose within the demonstration image. After achieving alignment, our framework transition to utilizing impedance control for full insertion of the object into the receptacle. We address the practical challenge of managing camera calibration errors and grasp pose disturbances simultaneously, a topic that has received limited attention in prior research. We adopt an object-centric approach to the visual servoing process and aim to reduce reliance on precise camera and grasp poses during computations. This strategy enhances the algorithm's robustness to uncertainties in these aspects.

In summary, we introduce a one-shot framework that seamlessly integrates 6-DoF visual tracking with a two-stage vision-force servo control for precise insertion of novel objects with just one static image demonstration of the task. Through comprehensive real-world experiments, we validate the effectiveness of our method across a variety of challenging insertion tasks. Our vision-force perception and servo system exhibits robust generalization capabilities, accommodating variations in object poses, calibration discrepancies, and previously unseen shapes.

II. RELATED WORK

The problem of robotic assembly involving insertion, has been a focal point of research for many years [1]–[10], [25]. Several proposed solutions can be classified according to the sensory data they employ, encompassing force, images, tactile, and laser inputs. In force-based servo methods, a common strategy involves estimating hole position through contact state modeling [13], [26]. Planar search techniques [14] assess the positional relationship between the peg and the hole by analyzing the torque resulting from their positional disparity. These methods commonly employ approaches such as Archimedes spiral search, square spiral search, windmill search, and raster search (Fig. 2). However, this necessitates a dense and long search pattern, and restricts its applicability to scenarios where the peg and hole are already approximately aligned. Another avenue of research explores tactile feedback policies for industrial insertion tasks [25], albeit with restricted generalization capabilities.

Researchers have also examined visual servoing techniques for insertion tasks, with recent advancements focusing on combining deep learning with these methods. For example, deep keypoint extraction can improve visual servoing for intricate robotic manipulation tasks [27]. Although visual servoing presents advantages for accurate assembly, its precision is inherently affected by inaccuracies in perception and calibration. Prior studies have investigated a two-stage methodology incorporating both vision and force inputs for diverse applications [17], [19], [28]. But these methods either can only apply to specific object shapes and are sensitive to calibration errors. Pose estimation and tracking-based servoing methods are also employed for precise manipulation tasks [29]. Recent advancements in deep learning-based object pose estimation [30], [31] have gained traction for their ability to achieve better accuracy across various datasets. However, these methods often require extensive offline training for specific objects and categories. An alternative approach involves constructing a geometric model during tracking, as seen in recent model-free techniques [32]–[34]. Another recent advancement involves zero-shot model-based tracking, which compares current measurements against renderings from a provided CAD model of the object. Many of these methods are well-suited for tasks with broad workspace domains and flexible tolerances, such as those in mobile robotics and visual navigation. However, only a few studies have examined the robustness of pose estimation in high-precision tasks [29]. In this endeavor, we incorporate a zero-shot tracking approach based on iterative corresponding geometry [23] for precise insertion tasks, enabling generalization capabilities for our framework. Nonetheless, our framework remains versatile, not bound to any particular tracking methodology.

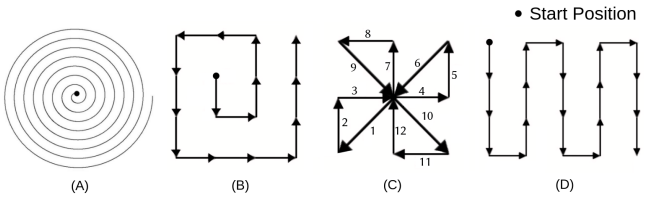

Fig. 2: Illustration of different search patterns [14] for insertion tasks: (A) Archimedes spiral search; (B) square spiral search; (C) windmill search; (D) raster search.

It is also noteworthy to mention the increasing trend of integrating reinforcement learning into insertion tasks [35], [36]. While these methods have shown considerable success, they often require complex learning setups, detailed reward engineering, a large number of demonstrations, and may struggle with generalizing to new objects, tasks, and unstructured environments. Our approach introduce a single-shot demonstration setting to mitigate the need for a training phase, while also reducing calibration errors and disturbances in in-hand object poses during the insertion process. The idea of demonstration-based insertion is inspired by the observation that humans do not require elaborate search patterns for insertion tasks [37]. Some studies employ kinesthetic teaching with force-feedback to gather demonstration data [38], while others utilize tracking systems to capture trajectories [21], [39] and replicate them for automated insertion. In contrast, our framework streamlines the process by requiring only a static image as demonstration of the task.

III. METHOD

In this section, we first introduce some preliminaries notations, we then formalize the insertion problem, followed by an explanation of our proposed **Insert-One** method.

A. Preliminaries

In this work, a coordinate transformation $\mathbf{T}_{SA} \in \mathbb{SE}(3)$ denotes a 4×4 homogeneous transformation matrix that describes the origin coordinates of a given frame $\{A\}$ and the orientation of its axes, relative to a given reference frame $\{S\}$. The transformation matrix represents a combination of rotation and translation in the 3D space, given by the rotation matrix $\mathbf{R}_{SA} \in \mathbb{SO}(3)$ and the translation vector $\mathbf{t}_{SA} \in \mathbb{R}^3$.

The transformation between the camera frame $\{C\}$, and a robot base frame of interest $\{B\}$ is denoted as $\mathbf{T}_{BC}$. This transformation can be computed using the standard procedure for base-eye calibration with a printed visual tag pattern. The base frame is fixed to the robot frame (i.e. center of the robot base). The tool pose refers to the end-effector frame at each time-step t, denoted as $\mathbf{T}_{BE}[t]$, that describes the position of the end-effector's origin and the orientation of its axes relative to the reference base frame $\{B\}$. This coordinate transformation can be automatically calculated using the joint angles and known kinematic equations. We define the distance between two transformations $\mathbf{T}_A$ and $\mathbf{T}_B$ as the sum of their rotation and translation distances:

$$\text{dist}(\mathbf{T}_A, \mathbf{T}_B) = \frac{\text{Tr}(\mathbf{R_A R_B}^T) - 1}{2} + \|\mathbf{t_A} - \mathbf{t_B}\|_2. \quad (1)$$

B. Problem Formulation

We consider the assembly-with-insertion tasks involving two objects. For simplicity, one object is designated as a stationary receptacle while the other serves as an insertion object that is randomly positioned on a flat surface at the start of the task. The robot is assumed to only know the pose of the receptacle. Initially, the robot must perceive and securely grasp the insertion object with its end-effector and then reposition it so that it snugly fits inside the receptacle. Unlike prior research, we eliminate certain assumptions. Specifically, our settings are unstructured in that we do not presume a rigid connection between the grasped object and the robot. Unlike previous methods where the object is firmly affixed to the robot's gripper [15] or other fixtures, we allow the object to translate and rotate during grasping and manipulation. The robot relies on vision feedback for in-hand object localization. The insertion process depends on tracking the grasped object's in-hand 6-DoF pose, estimated from color and depth images captured by RGB-D cameras. Furthermore, our goal is one-shot generalization with a single image demonstration, eliminating the need for a training phase for the insertion of novel objects.

C. Method Overview

To solve the previously outlined problem, we propose a two-stage vision-and-force servoing process that we refer to as **Insert-One** (Fig. 3). In the proposed mechanism, a tracking-based feedback control is used to achieve pose alignment while an adaptive search-based impedance control is used to execute the appropriate contact forces during the assembly. The first stage, that we call the *visual-alignment stage*, involves picking up the insertion object and then using a tracking-based feedback control to guide the object towards the pre-insertion pose. This pre-insertion pose is estimated from a single demonstration image of the task in our framework. A zero-shot model-based 6D tracker is deployed during this phase. The tracker returns the object's pose in the camera coordinates at each time-step. The object may not be perfectly centered in the end-effector after grasping. The visual alignment phase continues with visual feedback until the grasped object pose is sufficiently close to the pre-insertion pose.

After visual alignment, we transition to the second stage, that we call the *search stage*. It consists of a local search for performing an insertion into the receptacle with impedance control. The first stage is responsible for maneuvering the randomly presented insertion object accurately into a pose above the receptacle that is suitable for insertion. This first phase cannot ensure a successful insertion because of the high level of precision that is required, and the complexities associated with the contact-rich nature of the task. The search stage in our pipeline addresses and resolves these challenges.

D. Visual-alignment Stage

The objective of this stage is to execute a grasp that can effectively and dependably pick up the object from a randomly presented position, and then guide the object towards the pre-insertion pose above the receptacle.

6-DoF Object Pose Tracking: The object's 6-DoF pose is tracked to facilitate grasping, pre-insertion manipulation, and final assembly with continuous post-grasp displacement estimation. The challenge arises in industrial settings where many objects may vary vastly in shapes and textures, posing an extra training expense for object-level or category-level pose trackers. To address this and facilitate generalization, we employ a zero-shot model-based 6-DoF pose tracker, based on iterative corresponding geometry (ICG) [23]. This optimization-based tracker utilizes contour correspondences to iteratively refine the pose, making it adaptable to various novel objects without the need for pre-training in real-world scenarios. To address occlusion challenges during robotic assembly, we employ a dual-camera setup for effective pose tracking. We select one of the two cameras (camera 1 in Fig. 1) as the major camera. All operations involving camera coordinates are performed under the major camera's coordinates. Object pose $\mathbf{T}_{CO}[t]$ and end-effector pose in the camera's coordinates $\mathbf{T}_{CE}[t]$ are tracked at each time-step t, using the object and the robot hand CAD models.

One-Shot Image Demonstration: We provide a single image of the task to teach the robot where to position the object before insertion and how the object should be grasped. The object is moved into the pre-insertion pose right above the receptacle, through kinesthetic teaching (Fig. 3). An image I_g is taken from the major camera. Then, zero-shot pose tracking is used to estimate the pre-insertion pose of the

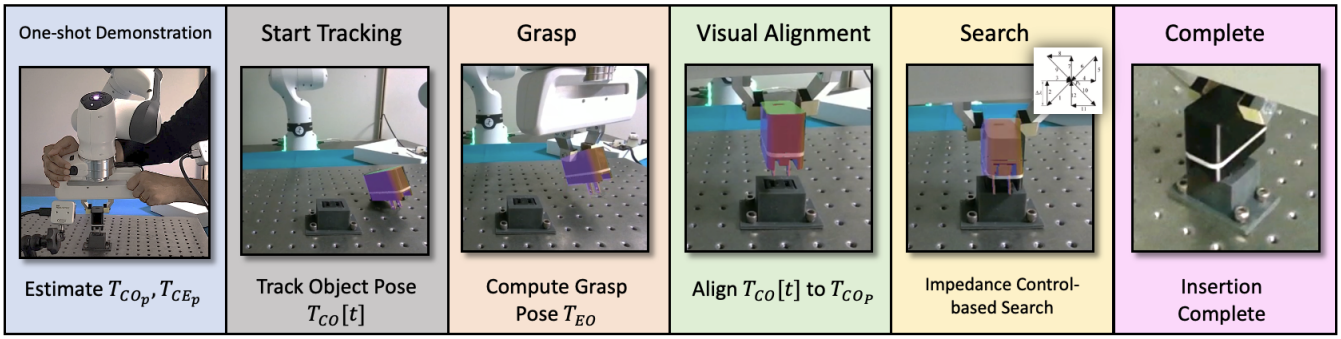

Fig. 3: Overview of Insert-One: Offline, we collect a single demonstration of a pre-insertion pose $\mathbf{T}_{CO_p}$ of the object. During testing, the object is placed in a random initial pose, and a 6-DoF real-time tracking is used to get the object pose $\mathbf{T}_{CO}[t]$ at each time t. Then, a object in-hand pose $\mathbf{T}_{EO}$ of the object in the end-effector's frame is computed and the object is grasped. During the visual alignment stage, a tracking-based feedback control is used by the robot to continuously synchronize the object's pose $\mathbf{T}_{CO}[t]$ with the demonstrated pose $\mathbf{T}_{CO_p}$. Finally, the robot uses an impedance control-based search along a windmill pattern to successfully insert the object.

object $\mathbf{T}_{CO_p}$ in the camera's coordinates. We also estimate the end-effector's pose from the image as $\mathbf{T}_{CE_p}$.

Grasping: At test time, the object is presented in a random pose on a flat surface. According to coordinate transformations, if we know both of the object's 6-DoF pose $\mathbf{T}_{CO}[t]$ and the robotic end-effector's pose $\mathbf{T}_{CE}[t]$ in the same camera coordinates, we can compute the object in-hand pose as:

$$\mathbf{T}_{EO}[t] = \mathbf{T}_{CO}[t] \cdot {\mathbf{T}_{CE}}^{-1}[t]. \tag{2}$$

The object in-hand pose $\mathbf{T}_{EO}[t]$ in Eq. 2 refers to the pose of the object in the end-effector's frame. From the demonstrated object pose $\mathbf{T}_{CO_p}$ and end-effector pose $\mathbf{T}_{CE_p}$, we compute an initial object in-hand pose $\mathbf{T}_{EO_p}$. The current end-effector's pose for grasping at t is then given as:

$$\mathbf{T}_{BE} = \mathbf{T}_{BC}\mathbf{T}_{CO}[t]\mathbf{T}_{EO_p}^{-1}. \tag{3}$$

We use position control to move the end-effector to $\mathbf{T}_{BE}$ and then conduct grasping. After grasping, we need to re-estimate the achieved object in-hand pose using Eq. 2 because sliding and other types of disturbance can occur during grasping.

Tracking-based Feedback Control: After picking up object O, we aim to move it to the demonstrated pre-insertion pose $\mathbf{T}_{CO_p}$. The control goal in this step is to minimize the error:

$$e_C = \text{dist}(\mathbf{T}_{CO}[t], \mathbf{T}_{CO_p}). \tag{4}$$

We know the object's pose $\mathbf{T}_{CO}[t]$ in the camera's coordinates from tracking. $\mathbf{T}_{BE}[t]$ is the pose of the robot's end-effector, which we can measure and control directly. According to the problem formulation, we also have a roughly calibrated camera pose $\mathbf{T}_{BC}$. By inserting the pre-insertion pose $\mathbf{T}_{CO_p}$ into this transformation chain, we can derive the desired end-effector pose at this stage as:

$$\mathbf{T}_{BE_p} = \mathbf{T}_{BC}\mathbf{T}_{CO_p}[t](\mathbf{T}_{EO}[t])^{-1}, \tag{5}$$

where $\mathbf{T}_{BE_p}$ is the estimated end-effector pose for pre-insertion. We use position control to move the end-effector to a reference position $\mathbf{T}_{BE_r}$. First, we set $\mathbf{T}_{BE_r} = \mathbf{T}_{BE_p}$. We name the control strategy until this step **Direct Control**.

In our framework, we follow a tracking-based feedback control for visual alignment. The reference control position, $\mathbf{T}_{BE_r}$ is updated in its rotation $\mathbf{R}_{BE_r}$ and translation $\mathbf{t}_{BE_r}$ separately using the following formula:

$$\mathbf{R}_{BE_r}[t+1] = \Phi(\mathbf{R}_{BC}\mathbf{R}_{CO_p}\mathbf{R}_{CO}^{-1}[t], K_R)\mathbf{R}_{BE}[t] \tag{6}$$

$$\mathbf{t}_{BE_r}[t+1] = K_t \cdot \mathbf{R}_{BC}(\mathbf{t}_{CO_p} - \mathbf{t}_{CO}[t]) + \mathbf{t}_{BE}[t] \tag{7}$$

Here, K_t and K_R are feedback gains for rotation and translation separately. $\Phi(\cdot, K_R) : \mathbb{SO}(3) \to \mathbb{SO}(3)$ is a mapping that scales the rotation by K_R.

$$\Phi(\mathbf{R}, K_R) = \exp(K_R \log(\mathbf{R})), \mathbf{R} \in \mathbb{SO}(3). \tag{8}$$

Here, exp and log are matrix exponential and logarithm. In practice, we set $K_t = 1$ and $K_R = 0.3$. Intuitively, we track the current pose of the object in the camera coordinates $\mathbf{T}_{CO}[t] = (\mathbf{R}_{CO}[t], \mathbf{t}_{CO}[t])$. Then, we compute the difference between $\mathbf{T}_{CO}[t]$ and the pre-insertion pose in the camera's coordinates $\mathbf{T}_{CO_p} = (\mathbf{R}_{CO_p}, \mathbf{t}_{CO_p})$. Since this residual is defined in the camera's coordinates, we transfer them back to the robot's frame using the robot-camera calibration $\mathbf{T}_{BC}$. We then apply this movement to the end-effector's current pose, $\mathbf{T}_{BE}[t] = (\mathbf{R}_{BE}[t], \mathbf{t}_{BE}[t])$.

A numerical simulation performed in Sec. IV-A shows that the visual alignment with feedback control is able to converge even in the presence of considerable errors in $\mathbf{T}_{BC}$ and $\mathbf{T}_{EO}$. The tracking stage converges when $e_C < \varepsilon$, where ε is a pre-defined parameter (set to $1e-5$).

E. Search Stage

Once the object has been moved to its pre-insertion position $\mathbf{T}_{CO_p}$, the fine-tuned search process commences. Although in the previous stage the robot has tried to move the object to $\mathbf{T}_{CO_p}$, which is right above the receptacle, a straightforward push down in practice cannot always succeed. There are two main reasons behind this failure: (1) The pre-insertion pose $\mathbf{T}_{CO_p}$ is estimated from the demonstration image I_g. Thus, $\mathbf{T}_{CO_p}$ includes a pose estimation error. (2)

The tracking-based feedback control has a control error. With these two combined, the initial position of the object at this stage has a small misalignment with the receptacle. The misalignment can be too small to be overcame by visual servoing alone. Therefore, instead of trying to predict the exact object's pose, the robot follows a search-based strategy for insertion with an impedance-control search method.

Task Space Impedance Control: A task space impedance controller (TSI) is used to make sure the object and the receptacle are being in contact during the search. The impedance controller formula can be written as follows:

$$F = M_d(\ddot{x}_d - \ddot{x}) + D_d(\dot{x}_d - \dot{x}) + K_d(x_d - x) \tag{9}$$

Here F is the force applied by the robot. $x, \dot{x}, \ddot{x}$ are the desired position, velocity, and acceleration in the task space. M_d, D_d, K_d are the desired mass, damping, and stiffness matrices. In practice, we set the damping and stiffness matrices as follows:

$$K_d = \begin{bmatrix} K_d^t & \mathbf{0} \\ \mathbf{0} & K_d^R \end{bmatrix}, \; K_d^t = 500 \cdot I_{3x3}, \; K_d^R = 100 \cdot I_{3x3}, \tag{10}$$

$$D_d = 2.0 \cdot \sqrt{K_d}. \tag{11}$$

Here, K_d^t is the stiffness for translation and K_d^R is the stiffness for rotation. We set the rotation stiffness K_d^R to be smaller than the translation stiffness K_d^t to make the end-effector's orientation easier to change so that the object and receptacle surface are in full contact. M_d is dependent on system identification. Here we use the default M_d in the Franka robot's official control library.

Search Strategy: There are many different search patterns as shown in Fig. 2. For our application, we select the windmill search pattern. This search strategy is chosen because the first stage of our approach has already roughly aligned the object to the center of the receptacle, so we need to select a symmetrical search pattern. In practice, if the object is cylindrical, one can also follow a spiral search. Here, we set the frame of reference's origin on the upper surface of the receptacle, with the z-axis pointing up. The goal value of the z coordinate for the controller is set to $1mm$ lower than the receptacle's upper surface in order to ensure that the object and the receptacle maintain contact during the search. In the meantime, the z-value of the end-effector's pose $\mathbf{T}_{BE}[t]$ is continuously monitored. Whenever a significant drop (larger than $2mm$) in that z-value is detected, the object is assumed to be already partially inserted. Subsequently, the target z-value of the control is reset to $5mm$ below the current z-value, which results in a full insertion.

IV. EXPERIMENTS

We conducted a series of experiments to validate our proposed framework's effectiveness. The goal was to assess the feasibility of addressing autonomous assembly challenges involving a variety of industrial insertion tasks with novel objects, while considering grasping and visual estimation errors. Our experimental framework encompasses both simulated scenarios and real-world robotic trials.

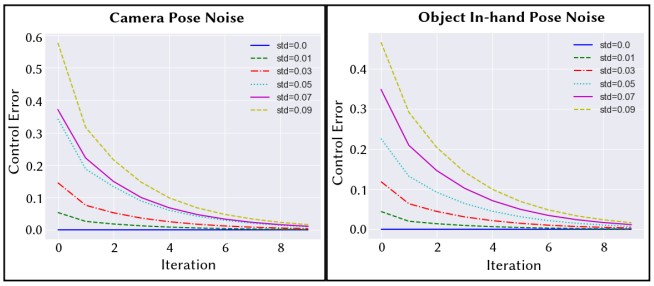

Fig. 4: Control error e_C as a function of iteration n, under different levels of injected camera pose noise (left) and object in-hand pose noise (right).

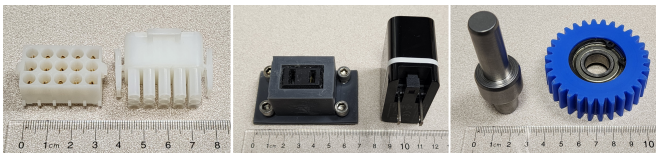

Fig. 5: Objects used in the experiments. From left to right, the objects are: AMP Connector PLUG (15-pin), AMP Connector HEADER, Power PLUG Adapter (NEMA 1-15P), Power PLUG Receptacle (NEMA 1-15R), Shaft (14.6 mm dia.), and Spur Gear (GEABDM2.0-30-20, 15 mm hole dia.).

A. Numerical Simulations

We performed numerical simulations for evaluating the convergence of our proposed visual alignment method under varying levels of injected object in-hand pose noise and camera pose noise. This provided a robustness analysis of our method. We generated the pose of insertion object in camera coordinates as,

$$\mathbf{T}_{CO}[t] = \mathbf{T}_{BC}^{-1}\mathbf{T}_{BE}[t]\mathbf{T}_{EO}. \tag{12}$$

We use ground-truth values for $\mathbf{T}_{BC}$ and $\mathbf{T}_{EO}$ to simulate the ground-truth $\mathbf{T}_{CO}[t]$ at each time-step. Then we perform the visual-alignment process as described in Sec. III-D numerically. The $\mathbf{T}_{BC}$ and $\mathbf{T}_{EO}$ we use during control is injected with Gaussian noise with variance σ, because we want to verify if the control can converge with existence of calibration error and object in-hand pose disturbance. We graphically represent the variation of the average control error e_C of 100 different goal poses as a function of control iterations n under different levels of object in-hand pose errors and camera pose errors, as illustrated in Fig. 4. As depicted in the results, both camera pose error and object in-hand pose estimation error significantly increase the initial control error e_C. However, with the progression of our tracking-based feedback control, we observe a gradual and steady reduction in this error.

B. Baseline & Ablation

Direct Control: Direct Control refers to a method that only use the estimated pre-insertion pose $\mathbf{T}_{BE_p}$ without the following tracking-based feedback control during the visual alignment stage. After end-effector is moved to $\mathbf{T}_{BE_p}$, we

follow the same impedance control search strategy as we use in Insert-One.

Insert-One w/o ICS: To investigate importance of the search stage, we test a variant of Insert-One without the impedance control search (ICS). After the first visual alignment stage, we only perform a push-down during the second stage.

C. Real-world Experiments

In this section, we describe the real-world experiments conducted to validate our methodology, in which we focus to assess the generalization, spatial invariance, and robustness of our approach.

Experimental Settings: The setup consists of a Franka Emika Panda robotic arm with 7 revolute joints and its control software running on a Desktop computer with Ubuntu 20.04. Two Intel RealSense D405 Depth Cameras are mounted for the multi-camera setup to facilitate 6-DoF visual tracking with RGB-D inputs. The experimental configuration is depicted in Fig. 1. The components featuring extruded characteristics (insertion objects) are positioned atop an optical breadboard and possess freedom of movement. Conversely, components with mating features (receptacles) are securely fastened to the breadboard to mimic industrial fixturing. The objective entails perceiving, tracking, grasping, transporting, and inserting all insertion objects into their respective receptacles. The tasks incorporate six distinct types of components, as illustrated in Fig. 5. These components vary in size, utility, number of pins, and visual appearance, offering a diverse range for assessment. The precision required for insertion tasks operates at sub-millimeter tolerances. Unlike prior studies that typically rigidly mount a test connector to the robot gripper or secure it in fixtures, our approach addresses the broader challenge of autonomous assembly with randomly positioned insertion object. This dynamic scenario accounts for handling uncertainty due to manipulation dynamics, with changes in object pose during grasping and manipulation.

Insertion Experiments: In the first set of experiments, distinct insertion tasks were conducted to test spatial invariance and generalization, wherein the insertion object was placed flat on a table in varying initial configurations. The tasks encompassed the insertion of a standard AMP 15-pin connector, an electrical adapter, and a Spur Gear (Misumi GEABDM2.0-30-20). Each configuration underwent testing from five distinct poses of the insertion object. For the first two objects, we set an offset angle θ set to 0°, 30°, 45°, -30°, or -45°. Since the spur gear is a symmetric object, we do not present offset in the initial gear pose, but instead, we manually introduce an in-hand perturbation to the gear pose after grasping, as shown in Fig. 6. The external perturbations, ranging approximately $\pm$1cm in translation and $\pm$10 degrees in rotation from the initial in-hand grasp. We conduct 10 trials for each task. An insertion is deemed successful only if the insertion object is fully seated in the receptacle. The outcomes of these experiments are summarized in Table I.

We compare our proposed method, Insert-One, against two alternatives: Direct Control combined with Impedance Control Search (ICS), and Insert-One without ICS. Results indicate that, although Direct Control and Insert-One without ICS each perform well in some trials, these methods are not stable to maintain performance across the variety of tasks, achieving 36.6% and 66.6% overall success rate, respectively. Direct control although accompanied by impedance search in the baseline, may not always minimize e_C sufficiently since the camera calibration $\mathbf{T}_{BC}$ is not accurate. The performance gain by our method (Insert-One) is significant, achieving overall 96.6% success. Ours consistently accomplishes the task regardless of the object category and initial conditions. The results highlights the importance of the visual alignment and impedance search module of Insert-One, for generalization and robustness in the proposed one-shot settings.

TABLE I: Results of insertion experiments on a variety of tasks. The test objects are shown in Fig. 5.

Task	Offset (°)	Method		
		Direct Control	Insert-One (w/o ICS)	Insert-One (Ours)
AMP Connector	0	2/2	2/2	**2/2**
	30	1/2	0/2	**2/2**
	45	2/2	0/2	**2/2**
	-30	2/2	0/2	**1/2**
	-45	2/2	0/2	**2/2**
Plug Adapter	0	1/2	2/2	**2/2**
	30	0/2	2/2	**2/2**
	45	0/2	2/2	**2/2**
	-30	0/2	2/2	**2/2**
	-45	1/2	2/2	**2/2**
Gear	-	0/10	8/10	**10/10**
Overall	-	11/30	20/30	**29/30**

TABLE II: Results of experiments on noise resistance. $\mathbf{T}_{EO}$ noise refers to object in-hand pose estimation noise. $\mathbf{T}_{BC}$ noise refers to camera pose noise. This experiment is performed with the AMP 15-pin Connector.

	Direct Control	Insert-One (w/o ICS)	Insert-One (Ours)
$\mathbf{T}_{EO}$ Noise	0/10	3/10	**7/10**
$\mathbf{T}_{BC}$ Noise	4/10	0/10	**9/10**

Robustness: We further investigated the noise resistance of our proposed approach in real-world conditions. As discussed in Sec III-D, the system inputs include the camera pose, $\mathbf{T}_{BC}$, and the object in-hand pose, $\mathbf{T}_{EO}$. To assess robustness in real-world settings, we manually introduced a translation noise with a standard deviation of 8*mm* and a rotation noise with a standard deviation of 8 degrees to $\mathbf{T}_{BC}$ and $\mathbf{T}_{EO}$ separately in two experiments. Subsequently, we ran our proposed system and other baseline methods under these disturbances. The results, as detailed in Table II, demonstrate that our proposed Insert-One framework effectively resists disturbances, unlike the baseline methods.

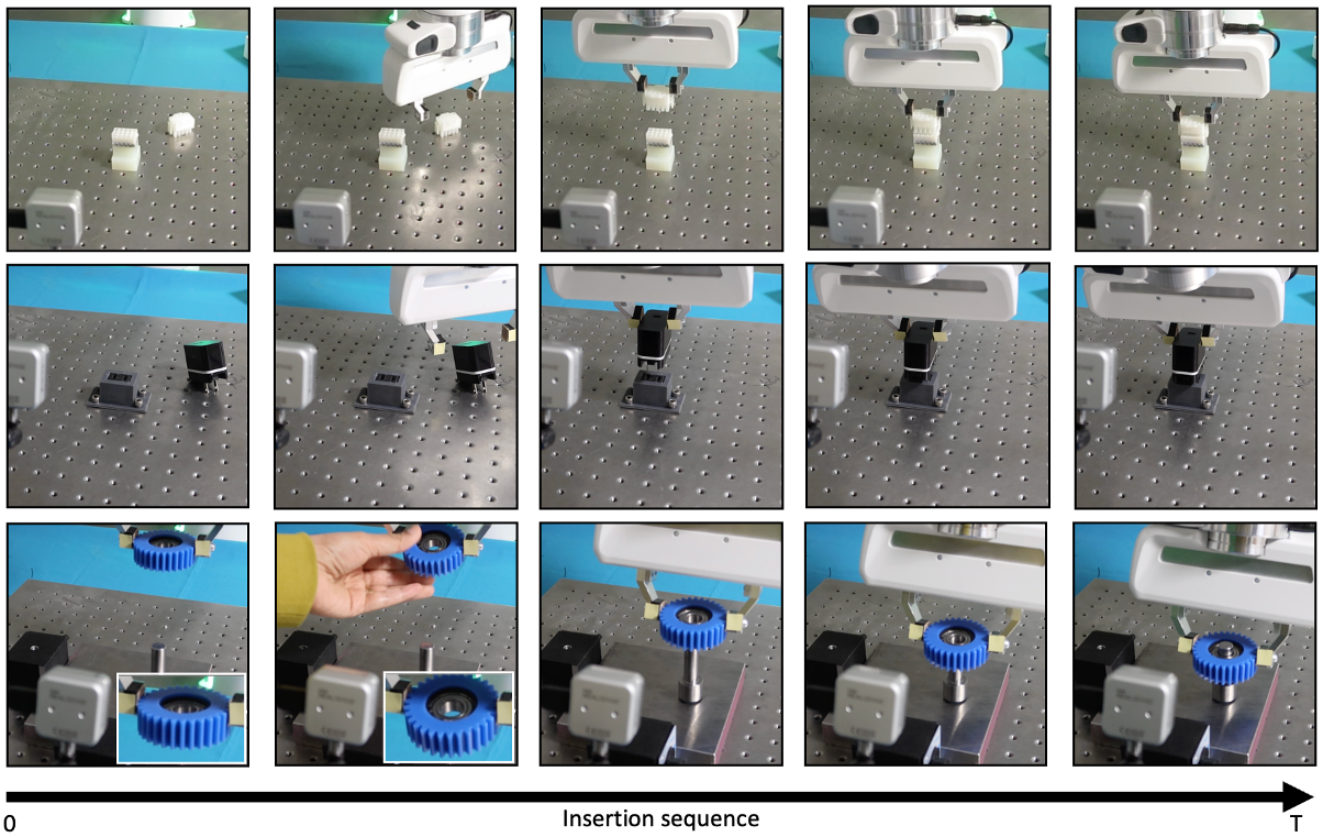

Fig. 6: Real-world experiments. From top to bottom, the inserted objects are: (1) AMP connector, (2) Plug Adapter, and (3) Gear. We showcase five key frames from each experiment: (1) beginning of tracking, (2) grasping, (3) visual alignment stage, (4) search stage, and (5) insertion completion. The gear insertion experiment evaluated robustness, wherein the gear pose is disturbed externally after its grasped by the robot. The two zoomed insert pictures show the difference between the gear's poses before and after external disturbance. Insert-One effectively withstands disturbances and demonstrates successful performance in insertion tasks.

D. Remarks

The major strength of our proposed framework lies in that using only a single image demonstration of the task, it excels in inserting new objects with high accuracy in precise manipulation tasks. Moreover, it showcases resilience to uncertainties stemming from calibration and disturbances in object pose, rendering it highly effective for practical applications. We don't compare to learning-based approaches, as our focus lies in generalization within one-shot settings, devoid of any pre-training on objects and tasks. While we show the effectiveness of the proposed method for insertion tasks, it can also be readily adapted for other applications, including object re-orientation. Relaxing certain assumptions outlined in this study could open up avenues for future research directions. While achieving zero-shot 6-DoF pose tracking necessitates a clear understanding of the geometry of the tracked objects, and to some extent their distinction from the background, this becomes notably more intricate with smaller objects. Although geometric information is typically accessible for numerous industrial parts, it highlights a limitation that could be mitigated through the development of more advanced tracking methods. Nevertheless, integrating additional sensor modalities, such as tactile sensing, along with improved search strategies based on contact information, holds potential for advancements.

V. CONCLUSION

We have presented Insert-One — a novel hybrid framework designed for the robotic insertion of new objects in a single-shot scenario, utilizing just a single image demonstration of the task. It integrates 6-DoF visual tracking-based iterative control and impedance control to facilitate the insertion of objects. Extensive experiments conducted across various high-precision assembly tasks highlight the effectiveness of the proposed framework. Demonstrating resilience against uncertainties stemming from camera pose calibration errors and disturbances in object in-hand pose, this framework enables seamless insertion of new objects—from grasping to transport and insertion into a receptacle with just one demonstration, unlike existing methods that typically require extensive training in controlled environments. Future efforts will focus on multi-modal sensory feedback and extending the method to cover multi-stage tasks.

REFERENCES

[1] I. C. Baykal, "Inspection of screw holes on machine parts using robot vision," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, pp. 1–5.

[2] W.-C. Chang, Y.-H. Weng, Y.-H. Tsai, and C.-L. Chang, "Automatic robot assembly with eye-in-hand stereo vision," in *2011 9th World Congress on Intelligent Control and Automation*, 2011, pp. 914–919.

[3] J. Jiang, L. Yao, Z. Huang, G. Yu, L. Wang, and Z. Bi, "The state of the art of search strategies in robotic assembly," *Journal of Industrial Information Integration*, vol. 26, p. 100259, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X21000571

[4] B. Mei, W. Zhu, G. Yan, and Y. Ke, "A new elliptic contour extraction method for reference hole detection in robotic drilling," *Pattern Analysis and Applications*, vol. 18, pp. 695–712, 2015.

[5] C. Sun, Q. Liu, and C. Wu, "Technique of log piles location based on visual saliency in port," in *2016 International Conference on Industrial Informatics-Computing Technology, Intelligent Technology, Industrial Information Integration (ICIICII)*, 2016, pp. 105–108.

[6] J. Wang and H. Cho, "Micropeg and hole alignment using image moments based visual servoing method," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 3, pp. 1286–1294, 2008.

[7] F. Xu, S. Wang, and B. Li, "Industrial robot base assembly based on improved hough transform of circle detection algorithm," in *Proceeding of the 11th World Congress on Intelligent Control and Automation*, 2014, pp. 2446–2450.

[8] Z. Yang, W. Liu, H. Li, and Z. Li, "A coaxial vision assembly algorithm for un-centripetal holes on large-scale stereo workpiece using multiple-dof robot," in *2018 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2018, pp. 1–6.

[9] Z. Zhang, J. Zhang, and D. Xu, "Design of microassembly system and research on coarse-to-fine alignment strategy in combination with active zooming," in *2013 IEEE Workshop on Robot Vision (WORV)*, 2013, pp. 76–81.

[10] P. Nagarajan, S. Saravana Perumaal, and B. Yogameena, "Vision based pose estimation of multiple peg-in-hole for robotic assembly," in *Computer Vision, Graphics, and Image Processing: ICVGIP 2016 Satellite Workshops, WCVA, DAR, and MedImage, Guwahati, India, December 19, 2016 Revised Selected Papers*. Springer, 2017, pp. 50–62.

[11] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1395–1476, 2021.

[12] I. F. Jasim, P. W. Plapper, and H. Voos, "Position identification in force-guided robotic peg-in-hole assembly tasks," *Procedia Cirp*, vol. 23, pp. 217–222, 2014.

[13] I. F. Jasim and P. W. Plapper, "Contact-state recognition of compliant motion robots using expectation maximization-based gaussian mixtures," in *ISR/Robotik 2014; 41st International Symposium on Robotics*, 2014, pp. 1–8.

[14] J. Jiang, L. Yao, Z. Huang, G. Yu, L. Wang, and Z. Bi, "The state of the art of search strategies in robotic assembly," *Journal of Industrial Information Integration*, vol. 26, p. 100259, 2022.

[15] G. Schoettler, A. Nair, J. Luo, S. Bahl, J. A. Ojea, E. Solowjow, and S. Levine, "Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5548–5555.

[16] Y. Zheng, X. Zhang, Y. Chen, and Y. Huang, "Peg-in-hole assembly based on hybrid vision/force guidance and dual-arm coordination," in *2017 IEEE international conference on robotics and biomimetics (ROBIO)*, 2017, pp. 418–423.

[17] H.-C. Song, Y.-L. Kim, and J.-B. Song, "Automated guidance of peg-in-hole assembly tasks for complex-shaped parts," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 4517–4522.

[18] Y.-l. Tang, Z.-j. Zhang, X. Ye, and X.-f. Zhang, "Micro-assembly precise coaxial alignment methodology based on surface roughness and reflectiveness matching," *Assembly Automation*, vol. 34, no. 2, pp. 141–150, 2014.

[19] H.-C. Song, M.-C. Kim, and J.-B. Song, "Usb assembly strategy based on visual servoing and impedance control," in *2015 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, 2015, pp. 114–117.

[20] O. Spector and D. Di Castro, "Insertionnet-a scalable solution for insertion," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5509–5516, 2021.

[21] B. Wen, W. Lian, K. Bekris, and S. Schaal, "You only demonstrate once: Category-level manipulation from single visual demonstration," *arXiv preprint arXiv:2201.12716*, 2022.

[22] G. Schoettler, A. Nair, J. A. Ojea, S. Levine, and E. Solowjow, "Meta-reinforcement learning for robotic industrial insertion tasks," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9728–9735.

[23] M. Stoiber, M. Sundermeyer, and R. Triebel, "Iterative corresponding geometry: Fusing region and depth for highly efficient 3d tracking of textureless objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6855–6865.

[24] 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," *arXiv preprint arXiv:2212.06870*, 2022.

[25] L. Fu, H. Huang, L. Berscheid, H. Li, K. Goldberg, and S. Chitta, "Safe self-supervised learning in real of visuo-tactile feedback policies for industrial insertion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 10 380–10 386.

[26] S. Jin, X. Zhu, C. Wang, and M. Tomizuka, "Contact pose identification for peg-in-hole assembly under uncertainties," in *2021 American Control Conference (ACC)*, 2021, pp. 48–53.

[27] E. Y. Puang, K. P. Tee, and W. Jing, "Kovis: Keypoint-based visual servoing with zero-shot sim-to-real transfer for robotics manipulation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 7527–7533.

[28] A. Stemmer, G. Schreiber, K. Arbter, and A. Albu-Schaffer, "Robust assembly of complex shaped planar parts using vision and force," in *2006 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2006, pp. 493–500.

[29] J. Liang, A. Boularias, A. Dollar, K. Bekris *et al.*, "Vision-driven compliant manipulation for reliable; high-precision assembly tasks," in *Proceedings of Robotics: Science and Systems*, 2021.

[30] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[31] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, "Dope: Distillation of part experts for whole-body 3d pose estimation in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 2020, pp. 380–397.

[32] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 8067–8074.

[33] H. Chang, D. M. Ramesh, S. Geng, Y. Gan, and A. Boularias, "Monostar: Mono-camera scene-level tracking and reconstruction," *arXiv preprint arXiv:2301.13244*, 2023.

[34] H. Chang and A. Boularias, "Scene-level tracking and reconstruction without object priors," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3785–3792.

[35] A. Nair, B. Zhu, G. Narayanan, E. Solowjow, and S. Levine, "Learning on the job: Self-rewarding offline-to-online finetuning for industrial insertion of novel connectors from vision," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[36] B. Tang, M. A. Lin, I. Akinola, A. Handa, G. S. Sukhatme, F. Ramos, D. Fox, and Y. Narang, "Industreal: Transferring contact-rich assembly tasks from simulation to reality," *arXiv preprint arXiv:2305.17110*, 2023.

[37] E. Klingbeil, S. Menon, and O. Khatib, "Experimental analysis of human control strategies in contact manipulation tasks," in *2016 International Symposium on Experimental Robotics*. Springer, 2017, pp. 275–286.

[38] J. Song, Q. Chen, and Z. Li, "A peg-in-hole robot assembly system based on gauss mixture model," *Robotics and Computer-Integrated Manufacturing*, vol. 67, p. 101996, 2021.

[39] Y. Yang, L. Lin, Y. Song, B. Nemec, A. Ude, A. G. Buch, N. Krüger, and T. R. Savarimuthu, "Fast programming of peg-in-hole actions by human demonstration," in *2014 International Conference on Mechatronics and Control (ICMC)*, 2014, pp. 990–995.