2024 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)
July 15-19, 2024. Boston, MA, USA

Visuo-Tactile Keypoint Correspondences for Object Manipulation

Jeong-Jung Kim[1], Doo-Yeol Koh[1] and Chang-Hyun Kim[1]

***Abstract*— This paper presents a novel manipulation strategy that uses keypoint correspondences extracted from visuo-tactile sensor images to facilitate precise object manipulation. Our approach uses the visuo-tactile feedback to guide the robot's actions for accurate object grasping and placement, eliminating the need for post-grasp adjustments and extensive training. This method provides an improvement in deployment efficiency, addressing the challenges of manipulation tasks in environments where object locations are not predefined.**

We validate the effectiveness of our strategy through experiments demonstrating the extraction of keypoint correspondences and their application to real-world tasks such as block alignment and gear insertion, which require millimeter-level precision. The results show an average error margin significantly lower than that of traditional vision-based methods, which is sufficient to achieve the target tasks.

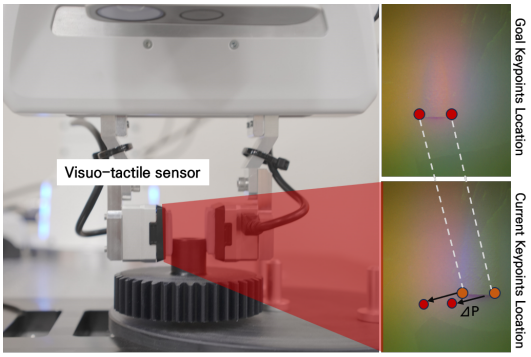

Fig. 1. Displacement estimation based on keypoint correspondences from visuo-tactile sensor images for pose adjustment in robot manipulation

I. INTRODUCTION

In the field of robotics, manipulation tasks that focus on the precise picking and placing of objects pose significant challenges, especially in environments where object locations are not predefined. Achieving precise manipulation in these environments requires advanced perception capabilities that allow robots to adapt their actions according to the identified object pose.

Traditionally, vision-based methods with color and depth cameras or LiDAR sensors have been used to estimate the pose of objects. However, these methods are often susceptible to sensor noise and environmental disturbances, which can affect the accuracy of manipulations. To overcome these limitations, research has explored the fusion of different sensing methods, including the novel application of visuo-tactile sensors. These sensors typically utilize a flexible elastomer material and a color sensor [1], [2], [3], [4]. It enables the transformation of tactile data into visual images. These sensor components are attached to the robot's end effector or the tip of a gripper. This configuration allows for direct observation of the object's contact state during manipulation tasks.

Compared to traditional visual-based methods, such tactile sensors provide improved manipulation accuracy. The direct sensing of contact state through tactile feedback enhances the robot's ability to grasp and manipulate objects with greater precision. This visuo-tactile sensors have been applied to various tasks including grasping, part identification, pose refinement and stability assessment during manipulation, although challenges such as the need for extensive training [5] and object marking [6] for improved recognition remain.

This paper presents a novel manipulation approach that uses keypoint correspondences from images captured by a visuo-tactile sensor to guide manipulation. After extracting feature descriptors from both the goal image and the current acquired sensor image, we compare the values corresponding to predefined keypoints in the feature descriptors of the goal image with the similarities between the entire feature descriptors in the acquired image. Following this comparison, we proceed to select the point with the highest similarity for finding correspondences. We conduct displacement estimation based on the keypoint correspondences and pose adjustment for robot manipulation. This approach has two advantages: it eliminates the need for additional adjustments after grasping and eliminates the requirement for extensive training, making deployment more efficient and faster.

The research has two contributions. Firstly, we propose a method that uses keypoint correspondences from visuo-tactile sensor data to enable precise manipulation without the need for additional learning. Secondly, we demonstrate the feasibility of this approach in real-world tasks, showing its effectiveness and reliability in enhancing manipulation precision.

The paper is organized as follows: Section II provides a detailed explanation of the proposed method, explaining the technical aspects and underlying principles. Section III presents the experimental setup and results, validating the efficacy of our approach across manipulation scenarios. Finally, Section IV concludes the paper by discussing the proposed approaches and their significance for manipulation,

* This study is a part of the research project, "Development of core technologies for robot general purpose task artificial intelligence (RoGeTA) framework (NK248G)", which has been supported by a grant from National Research Council of Science & Technology under the R&D Program of Ministry of Science.

[1] The authors are with Department of AI Machinery, Korea Institute of Machinery & Materials, Daejeon, Korea {rightcore, dyk, chkim78}@kimm.re.kr

as well as suggesting future research directions.

II. MANIPLUATION WITH TACTILE KEYPOINTS CORRESPONDENCES

This paper presents a novel framework for manipulation using keypoints extracted from images captured by visuo-tactile sensors. Keypoint correspondences have previously been shown to be effective for object pose estimation and adaptable across variations within object categories [7], [8]. Our research expands on this approach by utilising it for visuo-tactile sensor data, allowing for accurate manipulation tasks through focused interaction with the object of interest.

Visuo-tactile sensors have the advantage of focusing on the object in contact, which facilitates accurate position estimation. It also features self-illuminating components, enabling superior performance even in changing lighting environments. Additionally, our method utilizes foundation models, which are pre-trained deep learning models capable of understanding a wide range of data patterns and features. These foundation models offer several advantages, including their ability to identify features without requiring object-specific training. This not only simplifies the process but also enhances the system's adaptability across diverse tasks and domains, thereby increasing its versatility and usability.

Thus, our framework's applicability to a wide range of objects and tasks is greatly enhanced by this aspect, without the need for extensive learning phases for each new object category.

A. Overall Procedure

This section describes a two-phase manipulation process that uses visuo-tactile sensing and keypoint correspondence to achieve precise object handling. Our approach assumes that when an object is grasped, its features, such as points, lines, and textures, can be observed. These features can then be aligned between two images to establish correspondences. The proposed method is illustrated in Fig. 2 and detailed in Algorithm 1.

Correspondence points are identified between the sensor image acquired from human demonstration and the image captured during actual execution. These points are then used to calculate displacement, enabling pose adjustment based on the obtained values.

- Demonstration phase: In the first phase, a human demonstrates to the robot the grasp pose and grasping width required to perform a specific task. Although several approaches could be utilized for this purpose, our method involves the demonstrator manually positioning the robot's gripper to grasp the object. This action allows the visuo-tactile sensor to capture the object's shape, and keypoints are pre-defined and saved based on this captured image. In this process, we acquire a visuo-tactile image data $\mathbf{I}_g \in \mathbb{R}^{W\times H\times C}$ containing the target pose suitable for a task, K keypoints with (u, v) pixel coordinates $\mathbf{k}_g = \{u_i, v_i\}_{i=1}^{K}$ designated by the demonstrator, and the gripper width $w \in \mathbb{R}$.

Algorithm 1 Manipulation Algorithm with Visuo-Tactile Keypoints Correspondences

Initialize: Threshold for displacement τ
Demonstration Phase:
1) Human demonstrator positions the gripper and captures $\mathbf{I}_g$ stores gripper width w.
2) Define keypoints $\mathbf{k}_g = \{(u_i, v_i)\}_{i=1}^{K}$.
Execution Phase:
1) Attempt object grip and capture $\mathbf{I}_c$.
2) Process images through dense descriptor model:
    $f^D(\mathbf{I}_g)$ and $f^D(\mathbf{I}_c)$.
3) Apply correspondence function for finding correspondences between $\mathbf{k}_g$ and keypoints in $\mathbf{I}_c$:
    $\mathbf{k}_c = f^C(f^D(\mathbf{I}_g), f^D(\mathbf{I}_c), \mathbf{k}_g)$.
4) Estimate displacement:
    $\Delta\mathbf{P} = \text{EstimateDisplacement}(\mathbf{k}_g, \mathbf{k}_c)$.
if $|\Delta\mathbf{P}| < \tau$ **then**
    terminate with success.
else
    Adjust robot's end-effector pose in the Cartesian coordinate system by $\Delta\mathbf{P}$.
    Go to Step 1).
end if

- Execution phase: In the execution phase, the robot attempts to pick up the object using the approximate pose information obtained with a sensor such as a camera. It graps an object and then acquires a visuo-tactile image $\mathbf{I}_c \in \mathbb{R}^{W\times H\times C}$ from the sensor and identifies correspondences with the predefined keypoints $\mathbf{k}_g$.
  Tactile images $\mathbf{I}_g$ and $\mathbf{I}_c$ are processed through a dense descriptor model $f^D(\cdot)$, followed by a correspondence function $f^C(\cdot)$, resulting in $\mathbf{k}_c$, which corresponds to $\mathbf{k}_g$. Based on this correspondence, a displacement $\Delta\mathbf{P}$ is estimated, indicating the deviation from the target pose for manipulation in the Cartesian coordinate system.
  If the norm of the displacement $|\Delta\mathbf{P}|$ is below a predefined threshold, the process terminates; otherwise, the robot releases the object and adjusts its position by the calculated displacement amount, repeating the process as necessary.
  This iterative approach allows a more accurate determination of the grasp position and manipulation by taking into account how the object is held and adjusting accordingly. While a single attempt might fail due to recognition errors, iteration refines the robot's perception of the object's position and orientation. This approach provides a refined strategy for manipulation, combining visuo-tactile feedback with visual feature matching to improve the accuracy and reliability of manipulation.

B. Keypoints Correspondences

To build dense descriptors for the tactile sensor data at the step 2 of the execution phase, we used the DINO, which uses a pre-trained Vision Transformer (ViT) to extract deep

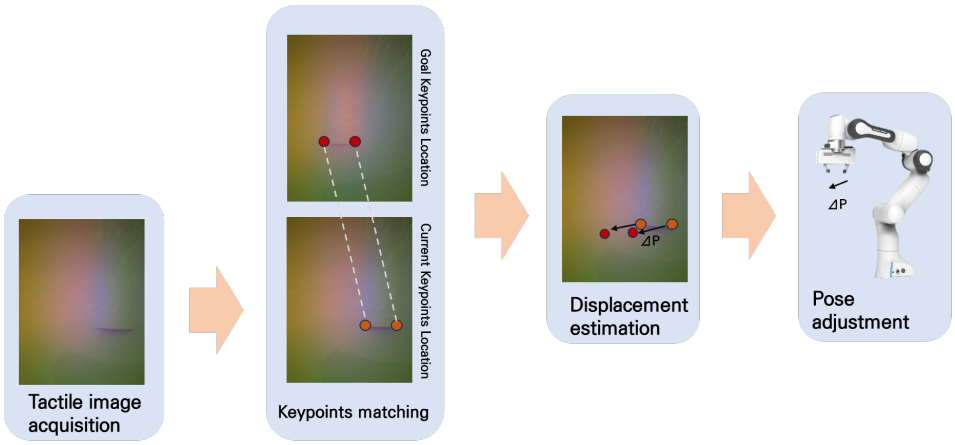

Fig. 2. Manipulation process using keypoints extracted from visuo-tactile sensor images. The correspondence points between the sensor image obtained from human demonstration and the image captured during actual execution are identified. Displacement is calculated using this correspondence, and a pose adjustment is performed based on the value.

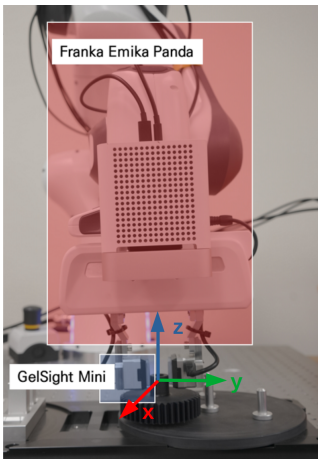

Fig. 3. Experimental setup. A GelSight Mini sensor, which is a visuo-tactile sensor, is attached to the end-effector of the Franka Emika Panda robot to acquire sensor data and estimate displacement.

features that serve as dense visual descriptors [9], [10]. These features capture strong, well-localised semantic information with a high degree of spatial granularity. Furthermore, the semantic information encoded in these features is applicable across a spectrum of related, yet distinct, object categories. In this paper, we used the DINO method to generate dense descriptors, but it is also possible to use other methods.

Depending on the characteristics of the object, more than one keypoint may be required. However, this paper focuses on using two keypoints to find correspondences, which is sufficient for calculating two-dimensional displacement in terms of position and angle. While three keypoints could allow for three-dimensional displacement calculations, the nature of visuo-tactile sensors limits the accuracy of depth measurements, making two-dimensional information more reliable for precise manipulation. However, since the number of keypoints required for pose estimation varies from object to object, it is necessary to adjust this parameter according to the specific problem at hand.

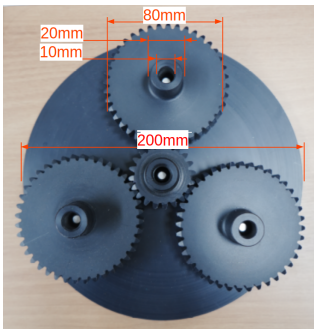

Fig. 4. Objects for gear insertion task. A robot picks up gears and inserts them into holes on a panel.

III. Experiments

To investigate whether keypoint correspondences can be extracted from visuo-tactile images, the precision of the displacement estimation method, and whether this method can be applied to real manipulation tasks, we conducted a series of experiments.

Our experimental setup consisted of equipping a Franka Emika Panda from Franka Robotics [1] with a GelSight Mini sensor from GelSight [2] at the gripper end of the robot, as shown in Fig. 3. This sensor captures contact information within an area of $18mm \times 24mm$ at a resolution of 240×320, which we adjusted to 224×298 for keypoint extraction. The experiment was conducted using only one of the two sensors attached to the robot. For feature extraction, we used the DINO method with the ViT-S/8 model, with a step size of 4.

The task performed by the robot and the size of the objects are shown in Fig. 4. This task involves picking up gears with holes and inserting them onto a shaft. The experiment was conducted with the gripper, equipped with sensors, grasping the upper part of the gear.

[1]Franka Robotics, http://www.franka.de/
[2]GelSight, http://www.gelsight.com/

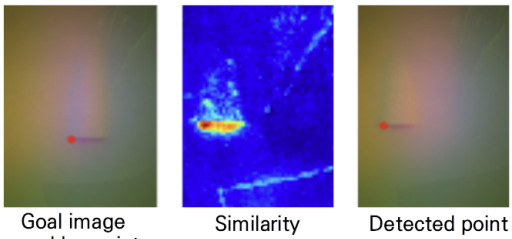

Fig. 5. Example of successful keypoint correspondence. Keypoint matching has been performed, associating the left corner of the object in the goal image with the left corner of the object in a captured tactile sensor data.

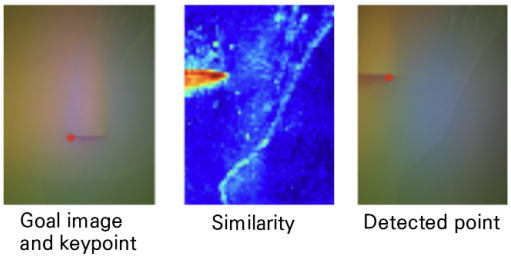

Fig. 6. Example of unsuccessful keypoint correspondence. The keypoint matching has incorrectly associated the left corner of the object in the goal image with the right corner of the object in a captured tactile sensor data.

A. Keypoint correspondences

The first experiment aimed to verify the effectiveness of extracting keypoint correspondences from images captured by the visuo-tactile sensor, and to measure the deviation of these keypoints from their ground truth positions. The experiments are conducted targeting the task of grasping the gear at the appropriate position to ensure successful insertion. The robot's end-effector was manually moved and oriented to the pose where insertion should occur. At this pose, an image acquired from the tactile sensor were set as the goal image, and keypoints were manually defined on the goal image.

We positioned the robot's end-effector at the pose and moved it randomly in a range of +/- 5mm in x and z axis to acquire test tactile images. The deviation between the keypoints extracted from these images and those identified by the operators was then evaluated. After extracting feature descritors using the DINO from both the goal image and the acquired sensor image, we compared the values corresponding to predefined keypoints in the descriptor of the goal image with the similarity between the entire descriptor in the acquired image. Following this comparison, we proceed to select the point with the highest similarity.

The displacement estimation error was calculated with (1),

$$d_{error} = \frac{\sum_{i=1}^{N} \sqrt{(p^i_{x_{real}} - p^i_{x_{est}})^2 + (p^i_{z_{real}} - p^i_{z_{est}})^2}}{N} \tag{1}$$

where $p^i_{x_{real}}$ and $p^i_{z_{real}}$ are real position of displacement which were measured from robot's end-effector position with kinematics information and $p^i_{x_{est}}$ and $p^i_{z_{est}}$ are estimated displacement calculated from suggested method, respectivley. We conducted 10 experiments, and the average error was 1.29 mm and its standard deviation was 0.71mm.

We found that in most cases the displacement estimation was succesed, as shown in Fig. 5. The left corner of the object in the goal image has been matched to the left corner of the object in the currently acquired image. However, when only part of the object was visible, or when it crossed the boundaries, keypoints corresponding to the opposite side were detected, leading to errors as shown in Fig. 6. The left corner of the object in the goal image has been incorrectly corresponded to the right corner of the object in the currently acquired image. One such case resulted in an error of 50 pixels, corresponding to an error of 3.75 mm. The limited detection range of this sensor leads to ambiguity at the boundary areas, resulting in such outcomes. However, this error margin, which is relatively small compared to vision-based methods, is considered sufficient to achieve the target tasks using techniques such as impedance control, assuming a rough alignment between features such as lines and points in the captured and target images.

B. Manipulation tasks

To demonstrate the capability of the proposed method for precise tasks, we conducted experiments on gear insertion and block alignment tasks, both of which required millimeter-level accuracy that could not be achieved with external cameras alone. The method enabled alignment followed by robot control via impedance control. The experiments successfully confirmed the feasibility of both tasks, as shown in Figs. 7 and 8. After gripping the object, the sensor image was acquired, and the displacement was estimated by performing keypoint correspondence. From the obtained correspondences, the displacement is estimated. Based on this estimated displacement, an offset adjustment is applied to the robot's end-effector pose, enabling it to re-grasp the object and complete the task. In cases of insufficient pre-alignment, the use of sensors such as force-torque sensors and iterative search techniques or reinfocement learning are necessary to achieve correct positioning [11]. However, our method reduces the burdens associated with completing the task, which is essential for applying the robot in real-world applications.

IV. CONCLUSION

In this paper, we have introduced a manipulation strategy that uses keypoint correspondences from visuo-tactile sensor images to improve the precision of object picking and placement tasks. This method not only reduces the need for post-grasp adjustments, but also minimises the dependency on extensive training, thus increasing deployment efficiency.

The experimental results have validated the effectiveness of our approach, demonstrating that keypoint correspondences can be accurately extracted from visuo-tactile images, with an average positional error low enough to allow precise manipulation through techniques such as impedance control. Furthermore, our method has proven capable of performing tasks requiring millimeter-level accuracy, such as

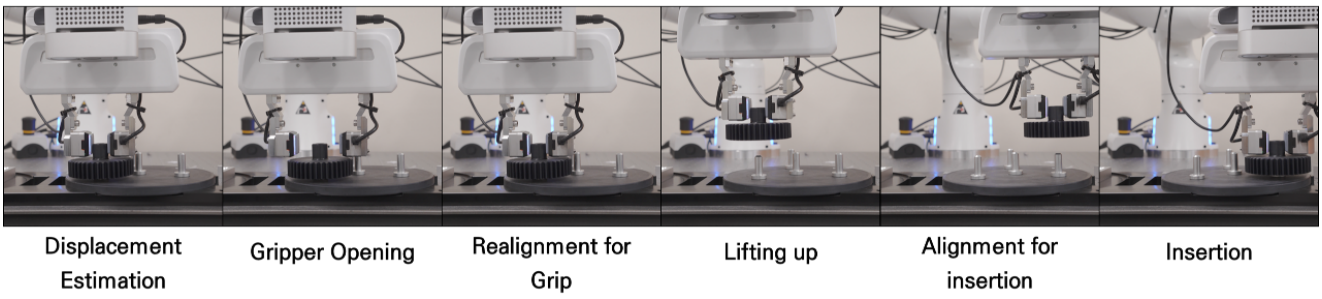

Fig. 7. Snapshot of gear insertion task using the proposed method

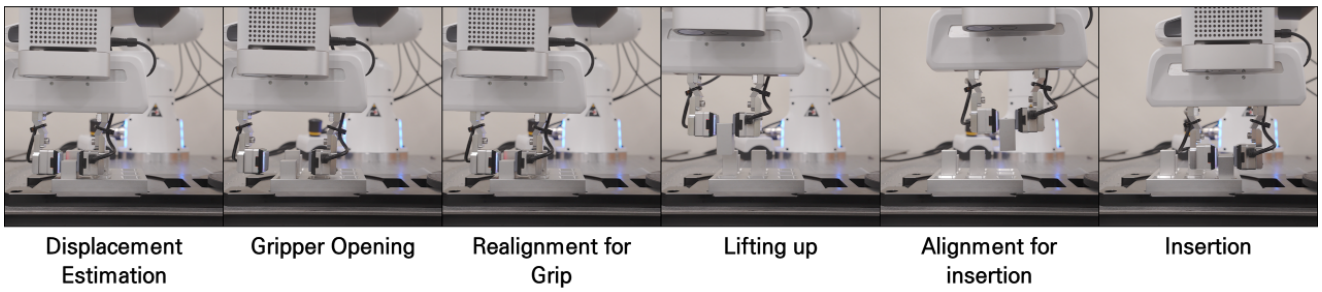

Fig. 8. Snapshot of block alignment task using the proposed method

block alignment and gear insertion, which are challenging for traditional vision-based systems.

However, our method requires a rough initial alignment and the presence of detectable features for successful keypoint extraction. This could be a limitation in scenarios where such conditions are not met, suggesting the need for further research into active alignment strategies. In addition, the current approach requires predefined keypoints for each object category, which could be a drawback when dealing with new categories.

Future research will address these limitations through the development of algorithms that can actively adjust the position of the robot's end effector for optimal feature extraction. Additionally, automating the process of keypoint selection for new object categories would increase the versatility and applicability of our method, enabling robots to perform more complex and varied tasks in dynamic and unstructured environments.

REFERENCES

[1] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[2] Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1927–1934. IEEE, 2018.

[3] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[4] Ian H Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787. IEEE, 2022.

[5] Siyuan Dong, Devesh K Jha, Diego Romeres, Sangwoon Kim, Daniel Nikovski, and Alberto Rodriguez. Tactile-rl for insertion: Generalization to objects of unknown geometry. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6437–6443. IEEE, 2021.

[6] Joyce Xin-Yan Lim and Quang-Cuong Pham. Grasping, part identification, and pose refinement in one shot with a tactile gripper. *arXiv preprint arXiv:2312.17650*, 2023.

[7] Peter Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2):492–499, 2019.

[8] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[10] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[11] Jianlan Luo, Eugen Solowjow, Chengtao Wen, Juan Aparicio Ojea, Alice M Agogino, Aviv Tamar, and Pieter Abbeel. Reinforcement learning on variable impedance controller for high-precision robotic assembly. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3080–3087. IEEE, 2019.