



# ActiveSPN: Active Soft Polyhedral Networks With Pose Estimation for In-Finger Object Manipulation

Sen Li , Chengxiao Dong, Chaoyang Song , *Senior Member, IEEE*, and Fang Wan , *Member, IEEE*

**Abstract**—Robotic grippers aim to replicate the remarkable functionalities of the human hand by providing advanced perception, adaptability, stability, and dexterity for complex tasks. Achieving these capabilities demands a sophisticated design hierarchy and robust perception mechanisms that ensure accurate manipulation. This letter introduces Active Soft Polyhedral Networks (ActiveSPN), a gripper design that leverages an active, non-biomimetic surface for precise in-hand manipulation. A vision system integrated directly into the fingers further facilitates accurate pose estimation of the in-finger object. The proposed system includes: (i) a soft polyhedral network featuring a transparent active belt to deliver complete three-dimensional adaptation and dexterous in-finger motion, and (ii) a generative learning-based pipeline for in-finger pose estimation. Experimental results demonstrate the ability of ActiveSPN to execute multi-degree-of-freedom in-finger manipulations, including two-axis rotation and one-axis translation. Moreover, the integrated vision-based pose estimation provides robust, real-time predictions, supporting consistent closed-loop control. Across diverse objects, the system achieves mean translational errors of 2.59 mm and rotational errors of 7°, highlighting a promising paradigm for compact, efficient, and dexterous robotic manipulation.

**Index Terms**—Vision-based deformable perception, in-finger manipulation, active surface.

## I. INTRODUCTION

**I**N-HAND manipulation refers to the ability to reposition or reorient an object within the hand or grasp without setting it down or releasing it. These include finger gaiting [1], leveraging external environmental contacts [2], palm supporting [3], throwing [4], and in-finger manipulation [5], [6]. Among these, in-finger object manipulation refers to the fine motor skill of manipulating an object using the fingers while it remains in contact with them. This type of manipulation generally involves

adjusting the object's position, orientation, or movement within the hand or fingers without the need to re-grasp the object from the environment. This distinctive feature allows continuous, localized in-finger manipulation to achieve dexterity in robotic systems. Unlike other strategies that rely on sequential grasping or external support, in-finger manipulation excels in tasks requiring precision and adaptability. This strategy is especially critical in highly dexterous applications, as it demands precise information about contact formation and stability to achieve robust and effective manipulation [7]. It is a critical skill for tasks that require precision and dexterity, which is fundamental in many everyday activities, from writing and drawing to handling small objects and tools [8].

Gripper design optimized for in-finger manipulation has attracted substantial research attention [3]. Anthropomorphic grippers are dexterous for their high degrees of freedom (DoFs) [9], but challenging in optimizing multi-DOF movements for robust manipulations, requiring comprehensive visual and tactile perception, as well as complex control algorithms. There are two main types of non-anthropomorphic approaches. One approach involves using underactuated mechanisms [10] to perform dexterous tasks while simplifying system structures by carefully selecting degrees of freedom (DoFs), which leads to decreased controllability. The other uses active surfaces [11], [12] to replace the long serial-chain DoFs, which allow continuous movement or rotation of an object. These active surfaces enable the gripper to move the hand-held object through locally imparted motions at the contact point. Innovative approaches to active surface mechanisms include movable components such as cylindrical rollers [13], belts [14], [15], and spherical rollers [5], which enable the continuous movement or rotation of an object. Beyond rigid structures, compliant materials and soft robotic designs [16], [17], [18], [19] have shown great potential in simultaneous grasping adaptability and perception capability. However, these soft fingers were mainly adaptive for grasping and lacked dexterity for more complex manipulation tasks. In contrast to other manipulation strategies, in-finger manipulation enables continuous and autonomous motion of the object along the finger's surface, minimizing reliance on environmental factors or complex re-grasping strategies. This capability facilitates seamless and uninterrupted operation, even in constrained or unstructured environments.

Despite its advantages, the complex mechanical designs of in-finger manipulation systems present challenges for integrating tactile sensors. Existing tactile sensing techniques, including photometric stereo-based sensors (e.g., GelSight) [20],

Received 14 February 2025; accepted 11 June 2025. Date of publication 26 June 2025; date of current version 2 July 2025. This article was recommended for publication by Associate Editor J. Qu and Editor C. Laschi upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grants 62206119 and Grant 62473189, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2025A151010424, and in part by the Shenzhen Long-Term Support for Higher Education at SUSTech under Grant 20231115141649002. (Corresponding author: Fang Wan.)

Sen Li, Chengxiao Dong, and Fang Wan are with the School of Design, Southern University of Science and Technology, Shenzhen 518055, China (e-mail: wanfang@iee.org).

Chaoyang Song is with the Design and Learning Research Group, Southern University of Science and Technology, Shenzhen 518055, China.

Codes are available at <https://github.com/ancorasir/ActiveSPN>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/LRA.2025.3583616>, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3583616

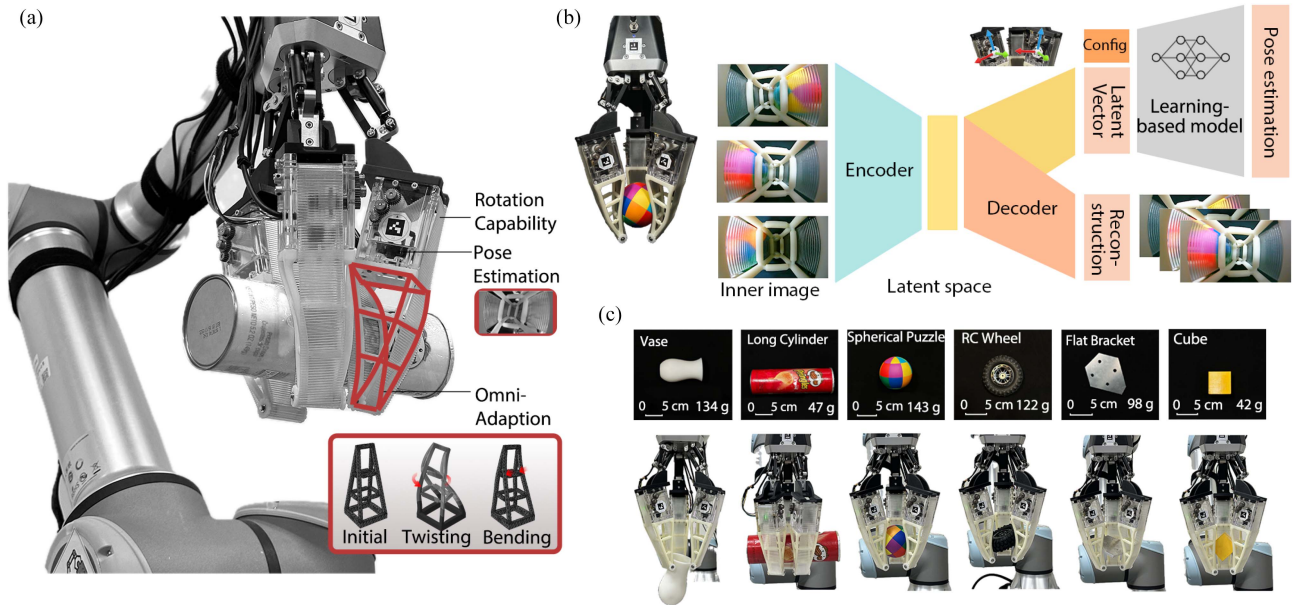


Fig. 1. Pose estimation of ActiveSPN with an omni-adaptive surface for in-finger manipulation. (a) The ActiveSPN finger with an active surface installed on existing grippers. (b) Overview of the ActiveSPN gripper architecture. (c) In-finger manipulation of diverse YCB objects via ActiveSPN.

barometric sensors [21], and resistance-, capacitance-, or piezoelectric-based sensors [22], [23], face significant limitations. Continuous rotation mechanisms make traditional wiring of sensors impractical, and photometric stereo systems often require fixed lighting rigs, restricting their application to controlled environments. Additionally, gel-based sensors are prone to damage under friction and pressure, further complicating their use on active surfaces. These challenges underscore the need for innovative solutions that maintain in-hand pose estimation without compromising the unique features of in-finger manipulation.

This study presents a novel ActiveSPN with transparent active surfaces, enabling in-hand manipulation with in-finger vision for pose estimation (Fig. 1) while leveraging active soft polyhedral networks for in-finger manipulation [24], [25] integrated with in-finger vision systems for tactile sensing [26]. This approach explores the potential of vision-based systems to perform object pose estimation and in-hand manipulation without relying on external aids. Using a three-finger gripper with ActiveSPNs, images captured by built-in cameras are processed through a learning-based model to extract object pose and finger configurations. The model employs a generative encoder-decoder architecture to predict and control the pose of the manipulated object. The main contributions are:

- *Design and fabrication of the ActiveSPN finger:* A robotic finger that integrates 3D omni-adaptability with an actively deformable surface, enabling in-finger manipulation through active surface motion and compliance.
- *Development of a semi-supervised variational autoencoder (SVAE) architecture for in-hand object pose estimation using in-finger vision:* This model predicts object poses and provides interpretable latent features to support downstream control and manipulation tasks.
- *Closed-loop in-finger manipulation demonstration:* The ActiveSPN gripper can estimate and manipulate object

pose in real-time tasks, closing the loop from robust perception to dexterous manipulation.

The following section describes ActiveSPN's kinematic model for in-finger manipulation. Section III presents our proposed in-finger manipulation method, with experiment results enclosed in Section IV. Final remarks are enclosed at the end.

## II. ACTIVESPN FOR IN-FINGER MANIPULATION WITH OMNI-DIRECTIONAL ADAPTATION

As shown in Fig. 2(a), the newly designed ActiveSPN finger comprises five main components: a soft polyhedral network element, a transparent timing belt, a built-in camera unit, a support frame, and a physical actuation system. The soft polyhedral network is the supporting framework, providing the necessary grasping force. Fabricated from polyurethane elastomer (Heicast 8400 from H&K) with a three-component ratio of 1:1:0 via vacuum molding, this component was chosen for its exceptional three-dimensional adaptability and its ability to fulfill passive omnidirectional adaptation requirements. The transparent belt is mounted by wrapping it around the surface of the soft polyhedral network, and its tension is adjusted using a timing belt pulley. The soft polyhedral finger weighs approximately 42 g and measures 125 mm in height.

The transparent timing belt, made of polyurethane elastomer with a hardness rating of 90 A and a thickness of 2 mm, serves as the active contact surface during manipulation. Transforming this component into a transparent belt facilitates direct observation of the grasped object. Its lightweight (21.7 g) and suitable stiffness and damping properties effectively increase the contact area at any given moment during manipulation. The recommended grasping force range is approximately 20 N, and the bending angle is preferably kept below 45° to ensure safe and stable operation. Excessive force can be generated, but

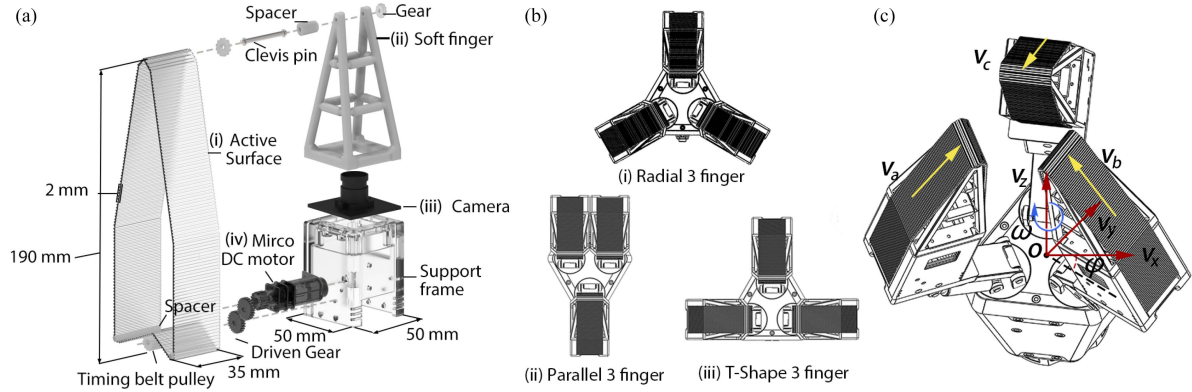


Fig. 2. ActiveSPN module's design and assembly. (a) Assembly of mechanical components of ActiveSPN, including (i) a transparent active surface, (ii) a soft polyhedral network, (iii) a camera with a support frame, and (iv) the motors and gearing mechanisms. (b) Reconfigurable grasping of a 3-finger robotic system, including (i) Radial, (ii) Parallel, and (iii) T-shape Configuration. (c) Kinematic model of radial 3-finger configuration.

may cause irreversible structural damage. Despite the concavity formed during finger curling, the belt remains stably in contact with the finger surface, as it is mechanically pressed between the finger and the object. The built-in camera unit provides real-time visual information about the object in contact during in-finger manipulation. In the current study, the ActiveSPN module employs a Chengyue WX605 camera (Weixinshijie) with a resolution of  $640 \times 360$  at 330 Hz.

Finally, the physical actuation system is designed for simplicity and minimal weight. Each ActiveSPN module used in the robotic fingers has an overall size of  $50 \times 50 \times 190$  mm and a weight of 196.8 g. The system mainly consists of a micro DC motor, a timing belt pulley, and gears. The micro DC motor drives the custom-designed timing belt to augment the soft finger backbone with an active surface, constituting the surface degree of freedom.

We introduce several improvements to our previous finger design [6], including the active surface with a transparent belt and the vision through a built-in camera, which allows for the observation of the object being held. Previous investigations [27], [28] have shown that the number and arrangement of fingers considerably influence the grasping outcome and overall robustness of the gripper. In this paper, we augmented a DH-3 gripper from DH-Robotics, a linkage-type adaptive gripper, and a servo-electric gripper by installing three activeSPN modules as extensions of its original rigid fingers.

As shown in Fig. 2(b), ActiveSPN has three configurations, including parallel 3-finger (P3), radial 3-finger (R3), and T-shape 3-finger (T3). Each configuration has three belts. Let  $b$  denote the body coordinate frame of the ActiveSPN.  $V_a^b$  is the velocity between the surface and the object at the contact point  $A$  of the body coordinate frame relative to the spatial frame, as viewed in the current body frame, which is given by  $V_{oa}^b = \text{Ad}_{g_{oa}} V_a^b$ . Similarly, the velocities transferred from points  $B$  and  $C$  to the object  $O$  are  $V_{ob}^b = \text{Ad}_{g_{ob}} V_b^b$  and  $V_{oc}^b = \text{Ad}_{g_{oc}} V_c^b$ , respectively.

The hypotheses for this simplified model are as follows: a) the ActiveSPN and the object being manipulated are in point contact; b) the surface friction of the ActiveSPN is sufficiently strong, resulting in no slide motion between the ActiveSPN and

the hand-held object. The grasped object's total velocity is

$$V_o^b = \text{Ad}_{g_{oa}} V_a^b + \text{Ad}_{g_{ob}} V_b^b + \text{Ad}_{g_{oc}} V_c^b, \quad (1)$$

where  $\text{Ad}_g : \mathbb{R}^6 \mapsto \mathbb{R}^6$  is an adjoint transformation of twists. A special case of the ActiveSPN is a radial 3-finger configuration, space  $120^\circ$  apart when designing the movement direction of each belt, as shown in Fig. 2(c). The system's kinematic equations are

$$\begin{bmatrix} v_x \\ v_y \\ \omega \end{bmatrix} = \begin{bmatrix} \sin \psi & \cos \psi & l \\ -\cos \psi & \sin \psi & l \\ 0 & -1 & l \end{bmatrix}^{-1} \cdot \begin{bmatrix} v_a \\ v_b \\ v_c \end{bmatrix}, \quad (2)$$

where  $v_x, v_y$  denote the linear velocities of the hand-held object in the  $\hat{x}_n, \hat{y}_n$  directions;  $\omega$  is the angular velocity;  $\psi$  is the angular configuration of the system;  $l$  represents the length or characteristic dimension related to the kinematic setup;  $v_a, v_b$  and  $v_c$  are the velocities of belts  $A, B$ , and  $C$ , respectively.

After defining the kinematics of the ActiveSPN, the control is ready to be interrogated. The pose of the handheld object can be controlled to target a specific location through kinematics, as proposed above. Once the ActiveSPN begins to manipulate the hand-held object, an automatic routine detects the pose configuration error  $\mathbf{g}_e$ , where  $\mathbf{g}_e = (p_e, R_e)$ , including the translation error  $p_e = \|p_c - p_d\|$  and the rotational matrix error  $R_e = R_d R_c^{-1}$ , to determine whether it should start the control procedure, where  $p_c$  and  $p_d$  are the current and desired positions of the object;  $R_c$  and  $R_d$  are the object's current and desired rotation matrices. In the control procedure, the control inputs are the reference pose  $\mathbf{g}_{\text{ref}}$ , current pose  $\mathbf{g}_c$ , and belt velocity. Consequently, when starting the control procedure, a PD controller was used to regulate the belt velocity.

### III. IN-FINGER POSE ESTIMATION FOR OBJECTS

1) *Encoder-Decoder Architecture*: A generative learning architecture is employed for in-hand object pose estimation, utilizing the encoder-decoder architecture, as shown in Fig. 3. The generative model incorporates a variational autoencoder that extracts latent features from real-time images captured by in-finger vision, which are assumed to follow given prior



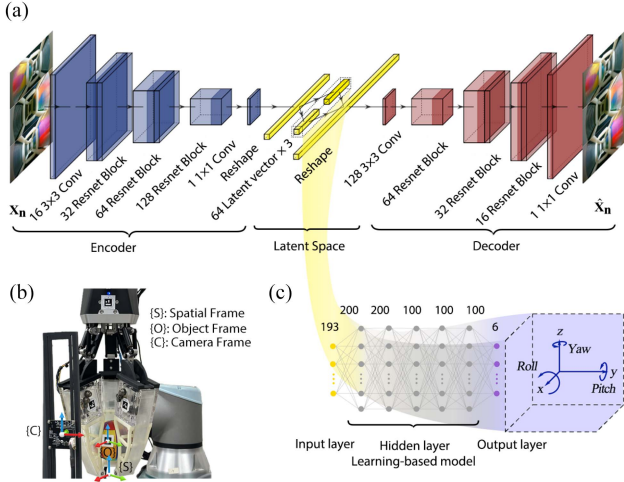


Fig. 3. Encoder-Decoder and Learning-Based Model Structures. (a) Encoder-Decoder architecture. (b) Platform for training data collection. (c) Auxiliary MLP for pose estimation.

normal distributions in the latent space. The encoder-decoder architecture is given by:

$$\begin{aligned} \mathbf{z}_n &= q_\phi^n(\mathbf{x}_n), \mathbf{n} \in \{1, 2, 3\}, \\ \hat{\mathbf{x}}_n &= p_\theta^n(\mathbf{z}_n), \mathbf{n} \in \{1, 2, 3\}, \end{aligned} \quad (3)$$

where  $\mathbf{x}_n$  represents the real-time image of the  $n$ -th finger,  $q_\phi$  is the probabilistic encoder resulting in a latent vector,  $\mathbf{z}_n$ ,  $\mathbf{z}_n = q_\phi^n(\mathbf{x}_n)$ ,  $p_\theta$  is the decoder function, and  $\hat{\mathbf{x}}_n$  is the reconstructed output image of the  $n$ -th finger. The distribution of the probabilistic encoder,  $q_\phi(\mathbf{z}_n|\mathbf{x}_n) \sim \mathcal{N}(\mathbf{z}_\mu(\mathbf{x}_n, \phi), \mathbf{z}_\sigma(\mathbf{x}_n, \phi))$ , and the decoder distributions  $p_\theta^n(\mathbf{z}_n) \sim \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$  followed a normal distribution.

The goal is to find an optimal representation,  $\mathbf{z}_n$ , of  $\mathbf{x}_n$  that encapsulates sufficient information about the object pose. To address the representation learning and pose estimation task, we use the VAE optimization framework [29] to incorporate an additional supervised task, maximizing the log-likelihood function of the marginal probability,  $\log p_\theta(\mathbf{x}_n)$ .

$$\begin{aligned} \log p_\theta^n(\mathbf{x}_n) &= \mathcal{L}(\theta, \phi; \mathbf{x}_n) + D_{KL}[q_\phi^n(\mathbf{z}_n|\mathbf{x}_n) || p_\theta^n(\mathbf{z}_n|\mathbf{x}_n)], \\ \mathbf{n} &\in \{1, 2, 3\}, \end{aligned} \quad (4)$$

where  $\mathcal{L}(\theta, \phi; \mathbf{x}_n)$  is the evidence lower bound (ELBO) for SVAE, which can be extended as follows:

$$\begin{aligned} \log p_\theta^n(\mathbf{x}_n) &\geq \mathcal{L}(\theta, \phi; \mathbf{x}_n) \\ &= \mathbb{E}_{q_\phi^n(\mathbf{z}_n|\mathbf{x}_n)}[\log p_\theta(\mathbf{x}_n|\mathbf{z}_n)] \\ &\quad - D_{KL}[q_\phi^n(\mathbf{z}_n|\mathbf{x}_n) || p_\theta^n(\mathbf{z}_n)], \end{aligned} \quad (5)$$

In (5), for continuous data such as images and latent variables  $\mathbf{x}_n$ , and  $\mathbf{z}_n$ , we assumed that the prior distribution of the latent variables,  $p_\theta^n(\mathbf{z}_n) \sim \mathcal{N}(0, \mathbf{I})$ , Maximization of the new ELBO in (5) was equivalent to maximizing the following optimization object, where the output of the decoder was denoted as  $\hat{\mathbf{x}}_n$ .

$$\tilde{\mathcal{L}}(\theta, \phi; \mathbf{x}_n) = \|\mathbf{x}_n - \hat{\mathbf{x}}_n\| - D_{KL}[\mathcal{N}(\mathbf{z}_n|\mathbf{x}_n) || p_\theta^n(\mathbf{z}_n)], \quad (6)$$

Therefore, we developed a hierarchical, convolutional, multiscale model for both the encoder and decoder to capture features from image data effectively. We used residual serial blocks to extract and reconstruct image features at different scales. The first terms in (6) measured reconstruction errors. The second term encouraged the approximated posterior  $q_\phi^n(\mathbf{z}_n|\mathbf{x}_n)$  to match the prior  $p_\theta(\mathbf{z}_n)$ , which controlled the capacity of latent information bottleneck. Although the derived optimization objective function (6) implicitly balances the three sources of loss, its optimization could be complex in practice. To resolve this issue, we proposed the formulation of (7) by introducing hyperparameters  $\omega$  and  $\alpha$  to (6). Note that  $\omega$  and  $\alpha$  are the parameters of the encoder-decoder neural networks, which need to be optimized during training.

$$(\omega, \alpha) = \arg \min_{\omega, \alpha} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|, \mathbf{n} \in \{1, 2, 3\}, \quad (7)$$

The encoder and decoder were trained with an Adam Optimizer with a learning rate of 0.001. The batch size was set to 32, the latent vector dimension to 64, and the number of epochs to 100. The dataset was split into a training set (12,000) and a validation set (3,000). Weights with the lowest validation loss were used for the experimental evaluation.

2) *Pose Estimation*: The auxiliary pose estimation model uses latent vectors of three fingers ( $N_1 = 64 \times 3$ ) and the gripper's configuration ( $N_2 = 1$ ) as input (Fig. 3(c)), which is formulated by

$$\begin{aligned} \mathbf{z}_n &= q_\phi^n(\mathbf{x}_n), \mathbf{n} \in \{1, 2, 3\} \\ \mathbf{g}_p &= f(\mathbf{z}_n, \mathbf{g}_{st}). \end{aligned} \quad (8)$$

where  $f$  is the regression model,  $\mathbf{g}_{st}$  is ActiveSPN configuration,  $\mathbf{g}_p$  is object's predicted pose. The regression model is a multilayer perceptron with five hidden layers, each containing 200, 200, 100, 100, and 100 neurons, respectively. In Fig. 3(b), we developed an experimental platform to collect training data. We trained an individual regressor head for each object category while freezing the encoder weights. The dataset for each object category was then randomly split into a training set of 4,000 samples and a validation set of 1,000 samples. The batch size was set to 32, and the optimization method was Adam, with a learning rate of 0.001. The best weights were obtained after 200 epochs, which were used as the final model. The pose consists of a rotation matrix and a translation vector. Thus, the loss function of the MLP is  $L = \sigma L_t + \gamma L_r$ , where the hyperparameters  $\sigma$  and  $\gamma$  are used to balance the rotational and translational loss. Considering the translation (in mm) and rotation (in radians) scales,  $\sigma$  and  $\gamma$  were set to 0.1 and 10, respectively.

3) *Data Collection Platform & Data Transformation in SE(3)*: The ActiveSPNs were installed on the fingertips of a rigid gripper (DH-3) and were then mounted on a robot (UR10e). ArUco tags were attached to the DH-3 gripper, the base frame of the ActiveSPN, and the objects to be manipulated to record the base position, which is used to calculate the configuration parameters of the finger. The ArUco tags acting on the object measure the ground truth pose of the manipulated object. Three in-finger cameras on the ActiveSPNs captured interaction deformation data, while an external camera recorded the poses of all markers

during data collection. The markers were  $4 \times 4$  squares with a width of 14 mm, which were detected by an external camera (1920  $\times$  1080) using OpenCV. The internal cameras were set to 640  $\times$  360 resolution, and the images were later resized to 320  $\times$  320 to reduce the model size and increase prediction speed. Data was continuously recorded while using a joystick to reorient the object. We simultaneously recorded the in-finger visual images and the corresponding pose labels.

The transparent surface allows in-finger vision to capture information on the texture and geometry of hand-held objects. Some studies [30] showed that geometry contributes to the in-hand pose estimation. However, further texture information is needed to distinguish poses of geometrically symmetrical objects. To evaluate ActiveSPN's performance in this aspect, we selected three symmetry objects from the YCB dataset: an irregular plastic peach, a spherical puzzle, and an irregular bottle cover, as test subjects, and collected 5,000 samples for each object. During data collection, we first set the gripper to force mode and controlled the gripper width to grasp the objects. The gripper kept still and remained as motionless as possible while the active surface movement drove the objects to collect pose data.

Post-processing of the transformation concerning the spatial coordinate system was performed after data collection. To describe the transformation from spatial frame  $S$  to object frame  $O$ , the object pose  $\mathbf{g}_{so}$  is  $(p_{so}, R_{so})$ , and the configuration space is the product space of  $R^3$  with  $SO(3)$ , which shall be denoted as  $SE(3)$  (for special Euclidean group),  $SE(3) = \{(p, R) : p \in R^3, R \in SO(3)\} = R^3 \times SO(3)$ . We denote the pose of frame  $O$  about frame  $S$  as  $\mathbf{g}_{so}$ .  $\mathbf{g}_{cs}$  and  $\mathbf{g}_{co}$  denotes the transformation from camera frame  $C$  to frame  $S$  and  $O$ , respectively. The desired poses are computed by  $\bar{\mathbf{g}}_{so} = \bar{\mathbf{g}}_{cs}^{-1} \cdot \bar{\mathbf{g}}_{co} \in SE(3)$ , thus,

$$\mathbf{g}_{so} = \begin{bmatrix} R_0^T R_{co} & R_0^T p_{co} - R_0^T p_0 \\ 0 & 1 \end{bmatrix}.$$

#### IV. EXPERIMENT AND RESULTS

##### A. Evaluation of Gripper Design on Grasping Success Rate

A total of 6 YCB objects were grasped, and each object was grasped 10 times by performing continuous grasping motions using different configurations of the ActiveSPN platform. The first contrast experiment evaluated how gripper configuration impacted grasping capability, specifically comparing three designs: R3, P3, and T3 (excluding the active surface). The second contrast experiment evaluated the effect of the active surface by comparing R3 and ActiveSPN R3. The third contrast experiment analyzed the role of in-finger manipulation, comparing the performance of ActiveSPN R3 with and without in-finger manipulation. In the manipulation condition, the active surface conveyed the hand-held object to the middle of the fingers before performing the power grasp. In contrast, the object was grasped directly without conveyance in the non-manipulation condition. We placed objects on a laboratory bench for each trial, and the UR10e robotic arm performed a grasping operation. We conducted grasping experiments with a fixed gripper width, testing each different object individually by placing them on

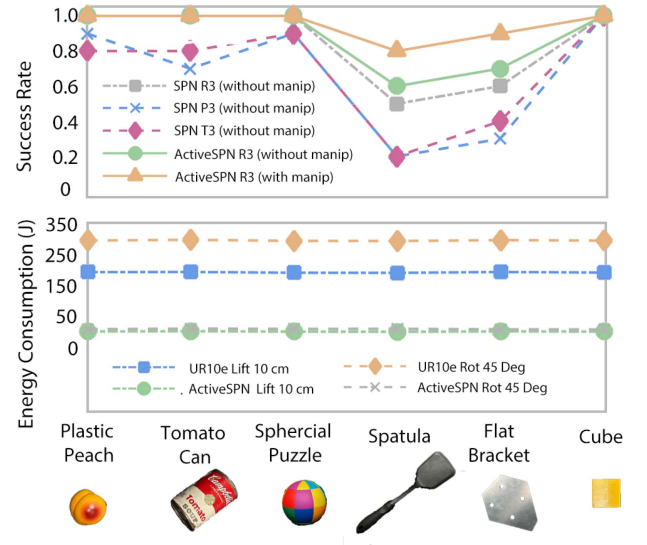


Fig. 4. Performance evaluation of grasp success rate and energy consumption for 6 YCB objects. (a) Grasp success rate of 6 YCB objects using T-shape, parallel, and radial configurations (without manipulation) and radial configuration with active and non-active surfaces. (b) Energy consumption of 6 YCB objects with different tasks (lift 10 cm and rotation 45°) using active surfaces and the UR10e robotic arm.

a laboratory bench for each trial, with the UR10e robotic arm performing the grasping operation.

The result in Fig. 4(a) helps to compare and contrast the results for the three experiments. The conclusions drawn from these experiments showed that: i) The radial 3-finger configuration can comparatively easily maintain stability. Compared to the parallel and T-shape 3-finger configuration, the radial structure provides a more uniform distribution of grasping forces, which improves stability during object manipulation. ii) The belts increase friction and improve the success rate of grasping. Adding belts to the design increases friction between the ActiveSPN and the object, thereby contributing to a higher success rate in grasping tasks. iii) The active surface uses control over the middle area of the ActiveSPN to improve the performance of grasping stability. Enabling the object to be conveyed from the fingertips to the middle area of the fingers is particularly well suited for power grasping. The result also indicates that the grasp success rate had no discernible differences among the finger configurations tested in this study. We conducted analyses of failed grasps and found that slim or thin objects, like spatulas and flat brackets, seem unsuitable for grasp because there is a 1-finger-width gap between two fingers. As such, we will implement the 3-finger radial configuration with in-finger manipulation in the following attempt.

##### B. In-Finger Manipulation of YCB Objects

Six sets of in-finger manipulation experiments were performed using YCB objects of various shapes and sizes. During the trial, the grasped object was oriented along a predetermined rotation axis, translated along a single axis, and oriented along a combination of two axes of rotation, typically not around a single axis. This procedure was developed by adapting the in-finger

competence of ActiveSPN, which implies the ability to carry out the operation in its entirety.

The iterative nature of our design process enabled us to expand the potential use and incorporate the in-finger manipulation capability into the prototype. Each finger of ActiveSPN has two actuated DoFs: a flexion DoF via DH-3 and a DoF via an active surface mechanism. The flexion DoF primarily provides force to make contact with the object, controlled via position control within the torque limits of the current servo of the DH-3. The active surface DoF's speeds were controlled using encoders coupled to the motor drive. To test whether the proposed robotic hand can skillfully manipulate various objects, we mounted the ActiveSPN onto a UR10e robotic arm and used ActiveSPN to grasp a series of YCB objects. As shown in Fig. 1(c), we demonstrated the ability to move different YCB objects with in-finger manipulation by ActiveSPN. Below each object, the ActiveSPN demonstrates in-finger manipulation for each shape, showcasing stability, dexterity, and adaptability to diverse geometries. Please refer to the supplementary video for further demonstration.

In practical industrial applications, robotic arms frequently need to lift or move objects, which typically requires significant energy consumption. The energy efficiency of these operations is a critical factor. To investigate this, we conducted experiments with the UR10e robotic arm and the ActiveSPN system. Each task was repeated five times to calculate the average energy consumption. As shown in Fig. 4(b), the UR10e robotic arm required approximately 2 s of joint actuation to lift an object by 10 cm, consuming an estimated 207.77 J of energy based on the power corresponding to the torque.

In comparison, the ActiveSPN system performed the same task with a significantly lower energy consumption of only 22.49 J. The UR10e robotic arm consumes approximately 307.58 J of energy to rotate an object by  $45^\circ$ , while the ActiveSPN system requires only about 30.12J. The UR10e robotic arm features six actuated joints, and adjusting the pose of an object at the end of the arm requires calculating the positions of all six joints for control, resulting in high energy consumption. However, in-finger manipulation within the ActiveSPN system eliminates the need for arm movement, leading to significant energy savings. In this operation, only the system's Micro DC motors, each with a torque of 0.33 Nm, are controlled to achieve the desired goal. Furthermore, within the load-carrying capacity of the ActiveSPN system, which is capable of manipulating objects ranging from approximately 1.5 to 10cm in size and weighing up to 0.85kg, using fewer motors and localized pose adjustments within the hand substantially reduces the energy required for manipulation compared to controlling the robotic arm's end effector. This enhanced energy efficiency is primarily achieved through the design of an optimized control strategy.

### C. Evaluation of In-Finger Object Pose Estimation

After finalizing data collection, the transformation analyses were conducted on the recorded dataset. We collected data on three objects: a spherical puzzle, an irregularly shaped bottle cap, and an irregularly shaped plastic peach. The pose distributions of the spherical puzzle are shown in Fig. 5(a). The figure shows

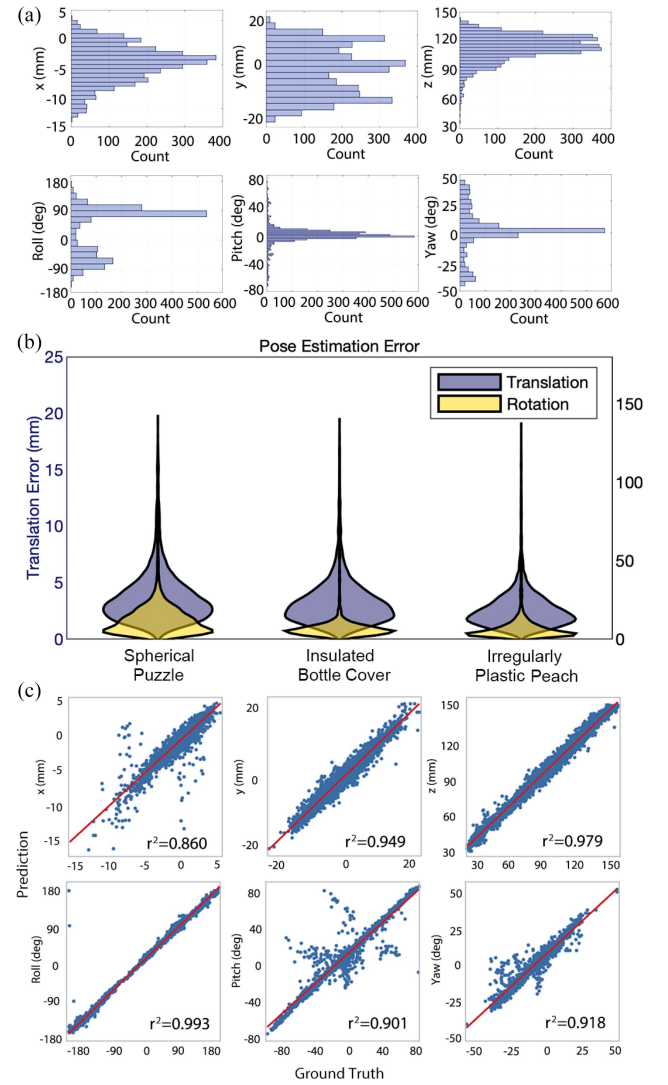


Fig. 5. Evaluation of pose estimation. (a) Distribution of (i) translations and (ii) orientations of the objects in the training dataset. (b) The pose estimation error histogram of three tested objects. (c) Distributions of prediction errors in each 6D pose dimension over different ranges.

that points are typically widely distributed along the  $Z$ -axis while centered around 0 mm on the  $x$  and  $y$  axes. The movement range of the object's marker is -50 to 150 mm along the  $Z$ -axis,  $\pm 20$  mm along the  $Y$ -axis, and almost aligned with the  $X$ -axis. The rotational distribution typically spreads over  $\pm 180^\circ$  in roll while centered around  $0^\circ$  in pitch and yaw. Because the ActiveSPN moves the hand-held object that translates in one direction and rotates around two axes, the movement oscillates  $\pm 180^\circ$  around the roll and  $\pm 80^\circ$  around the yaw, and the pitch is with very little motion. The pose estimated from each image was compared with the known pose (as determined by ArUco tags) to verify the accuracy of the estimation. Translation error is measured by the Euclidean distance between the predictive position  $t_{pre}$  and the ground truth position  $t_{gt}$ , denoted as  $t_e = \|t_{gt} - t_{pre}\|$ . Orientation error  $\theta_e$  is computed as the rotation matrix error  $R_e$ , if  $R_e = R_{pre}R_{gt}^{-1}$  is not an identity matrix, then,  $\theta_e = \cos^{-1} \frac{\text{trace}(R_{pre}R_{gt}^{-1}) - 1}{2}$ .



Fig. 5(b) shows the corresponding translation and rotation errors in pose estimation. We also plot Fig. 5(c) to show the predicted 6D pose via SVAE against the ground truth. The  $R^2$  scores are higher than 0.86 for 6D pose predictions, indicating the SVAE model's excellent performance in pose estimation on the test dataset. This estimation error histogram shows that their translation error distribution is more concentrated than the rotation error distribution. The irregular bottle cover object has an average translation error of 3.24 mm and an average rotation error of  $7.77^\circ$ . The spherical puzzle object has an average translation error of 3.68 mm and an average rotation error of  $12.55^\circ$ . The irregular plastic peach object has an average translation error of 2.59 mm and an average rotation error of  $7.00^\circ$ , suggesting better overall pose estimation accuracy than the other two objects. A comparison of the standard deviations shows that the irregular plastic peach object has a translation error standard deviation of 1.65 mm and a rotation error standard deviation of  $8.36^\circ$ , which is relatively concentrated, with minor fluctuations. These results indicate the method's precision and accuracy.

While multiple studies have investigated pose estimation for in-hand manipulation, none have incorporated active surfaces. The solution proposed in this study demonstrates excellent performance in translational error, with a range of 2.59–3.68 mm, which is comparable to the best results reported in [30] (2.02–4.00 mm) and [31] (2.14–7.31 mm) and significantly lower than those in [32] (15.00 mm) and [33] (7.66 mm). Regarding rotational angle error, this study achieved a range of  $7.00$ – $12.55^\circ$ , outperforming [30] ( $11.34$ – $31.87^\circ$ ), approaching the results of [34] ( $8.07^\circ$ ), and slightly exceeding [31] ( $0.81$ – $3.39^\circ$ ). Notably, the samples in [31] include minimal rotational scenarios, making their results less relevant for direct comparison in this context. This study achieves a well-balanced trade-off between translational and rotational angle errors, demonstrating high precision and stability. It thus provides a practical framework for designing in-finger manipulation grippers.

#### D. Closed-Loop In-Finger Manipulation

Closed-loop experiments were performed to validate the in-finger manipulation capability of the proposed ActiveSPN to estimate the pose and control the object to achieve the desired pose within the hand. We investigate the performance of pose measuring systems, which determine an object's pose by measuring a few fiducial markers (ArUco tag) attached to the object. The schematic diagram of ActiveSPN with a manipulated load controlled to a desired roll angle can be seen in Fig. 6(a). Fig. 6(b) shows sequential snapshots that capture the rotational control operation using ActiveSPN. Fig. 6(c) compares the actual and predicted values of the model and the input target angle of the controller of ActiveSPN during the in-finger manipulation experiment. The control system adjusted the roll angle to the target values, stabilizing the predicted angle below  $5^\circ$  with fluctuations under  $10^\circ$  during the movement.

Fig. 6(d) shows a schematic image of the hand-held object lifted during the in-finger manipulation using ActiveSPN. Fig. 6(e) shows sequential experimental snapshots tracking the

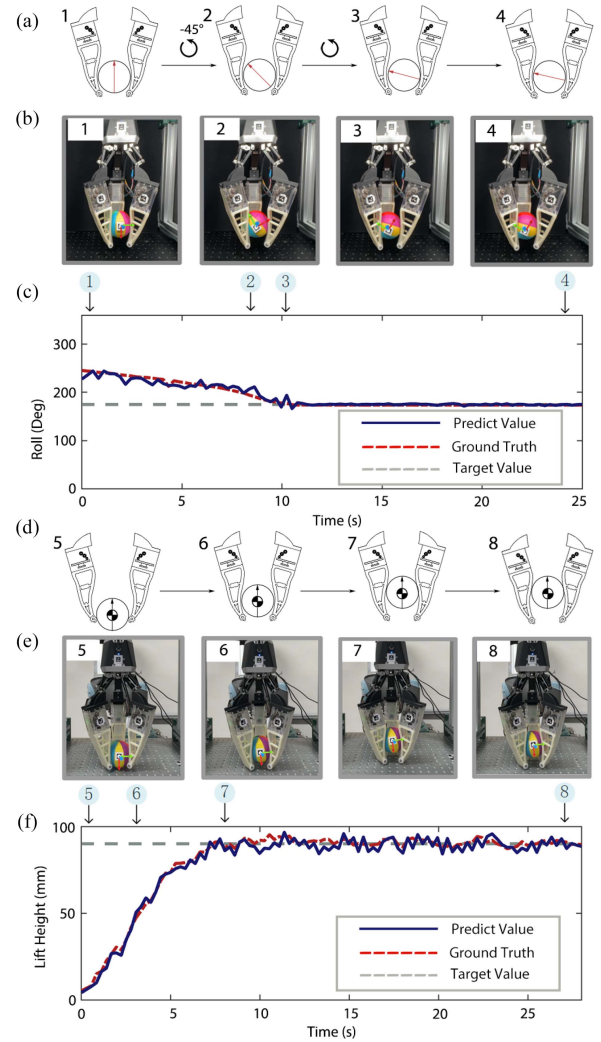


Fig. 6. Real-time object pose estimation using ActiveSPN for closed-loop in-finger manipulation. (a) Schematic illustration of rotation manipulation. (b) Snapshots during rotation manipulation. (c) Time series of rotation angle. The key timings are contact and start orientation at ①, orientation rotated by  $45^\circ$  at ②, orientation further rotated by  $20^\circ$  at ③ and stable placement at ④. (d) Schematic illustration of translation manipulation. (e) Snapshots during translation manipulation. (f) Time series of lift height. The key timings are contact and start translation at ⑤, translated to 45 mm at ⑥, and further translated to 90 mm at ⑦, and stable placement at ⑧.

pose of the object of the lifting process using ActiveSPN. The actual height attained by the outer camera was compared with the predicted height, and this showed a good correlation, as shown in Fig. 6(f). The target height of the object was set at 90 mm, and the actual result remained stable at around 90 mm, with fluctuations of less than 4 mm during the movement. Our experiments demonstrate the stable pose prediction and the control capability of the ActiveSPN controller.

#### V. CONCLUDING REMARKS

This study presents Active Soft Polyhedral Networks (ActiveSPN) as a novel robotic gripper that incorporates an active, transparent surface and integrated in-finger vision, enabling precise, omni-directional in-finger manipulation and robust pose estimation. Experimental evaluations confirmed that ActiveSPN

effectively utilizes its active surface combined with a semi-supervised generative learning model, achieving accurate real-time object pose estimation with mean errors of 2.59 mm in translation and 7 degrees in rotation. Additionally, ActiveSPN demonstrated superior energy efficiency compared to conventional robotic arm manipulation, significantly improving dexterity, adaptability, and operational efficiency.

Nonetheless, several limitations were identified in this study. Pose estimation accuracy may degrade for objects exhibiting high symmetry or similar textures, indicating a need for enhanced visual features or additional sensing modalities. Moreover, the reliance on manual data collection restricts scalability and adaptability, limiting the system's generalizability across broader object sets. Mechanical constraints, particularly the maximum recommended grasping force, further restrict the ability to handle fragile or extremely thin objects effectively.

Future research directions include enhancing the pose estimation framework by integrating supplementary tactile sensing or stereo vision systems to improve precision and robustness. Investigating automated or self-supervised approaches for acquiring training data will be crucial to scaling the system efficiently. Furthermore, refining the ActiveSPN design to accommodate a broader range of object sizes and delicate items, and rigorously evaluating its performance in dynamic and complex scenarios, will further validate its suitability for diverse real-world robotic manipulation applications.

## REFERENCES

- [1] O. M. Andrychowicz et al., "Learning dexterous in-hand manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3–20, 2020.
- [2] N. C. Daffle et al., "Extrinsic dexterity: In-hand manipulation with external forces," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2014, pp. 1578–1585.
- [3] C. Piazza, G. Grioli, M. G. Catalano, and A. Bicchi, "A century of robotic hands," *Annu. Rev. Control, Robot., Auton. Syst.*, vol. 2, no. 1, pp. 1–32, 2019.
- [4] K. M. Lynch and M. T. Mason, "Dynamic nonprehensile manipulation: Controllability, planning, and experiments," *Int. J. Robot. Res.*, vol. 18, no. 1, pp. 64–92, 1999.
- [5] S. Yuan et al., "Design and control of roller grasper V3 for in-hand manipulation," *IEEE Trans. Robot.*, vol. 40, pp. 4222–4234, 2024.
- [6] S. Li, F. Wan, and C. Song, "Active surface with passive omni-directional adaptation for in-hand manipulation," in *Proc. 6th Int. Conf. Reconfigurable Mechanisms Robots*, 2024, pp. 627–632.
- [7] J. Falco, K. Van Wyk, and E. Messina, "Performance metrics and test methods for robotic hands," NIST, Gaithersburg, MD, USA, Tech. Rep. 1227, 2018.
- [8] Z. Si, K. Zhang, O. Kroemer, and F. Z. Temel, "DeltaHands: A synergistic dexterous hand framework based on delta robots," *IEEE Robot. Automat. Lett.*, vol. 9, no. 2, pp. 1795–1802, Feb. 024.
- [9] M. Grebenstein et al., "The DLR hand arm system," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3175–3182.
- [10] R. R. Ma and A. M. Dollar, "An underactuated hand for efficient finger-gaiting-based dexterous manipulation," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2014, pp. 2214–2219.
- [11] V. Tincani et al., "Velvet fingers: A dexterous gripper with active surfaces," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2012, pp. 1257–1263.
- [12] K. Morino, S. Kikuchi, S. Chikagawa, M. Izumi, and T. Watanabe, "Sheet-based gripper featuring passive pull-in functionality for bin picking and for picking up thin flexible objects," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 2007–2014, Apr. 2020.
- [13] S. Yuan, A. D. Epps, J. B. Nowak, and J. K. Salisbury, "Design of a roller-based dexterous hand for object grasping and within-hand manipulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 8870–8876.
- [14] T. Nishimura and T. Watanabe, "Single-motor robotic gripper with three functional modes for grasping in confined spaces," *IEEE Robot. Automat. Lett.*, vol. 8, no. 11, pp. 7408–7415, Nov. 2023.
- [15] Y. Cai and S. Yuan, "In-hand manipulation in power grasp: Design of an adaptive robot hand with active surfaces," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 10296–10302.
- [16] F. Wan, X. Liu, N. Guo, X. Han, F. Tian, and C. Song, "Visual learning towards soft robot force control using a 3D metamaterial with differential stiffness," in *Proc. 5th Conf. Robot Learn.*, 2022, pp. 1269–1278.
- [17] X. Liu, X. Han, W. Hong, F. Wan, and C. Song, "Proprioceptive learning with soft polyhedral networks," *Int. J. Robot. Res.*, vol. 43, no. 12, pp. 1916–1935, 2024.
- [18] N. Guo et al., "Proprioceptive state estimation for amphibious tactile sensing," *IEEE Trans. Robot.*, vol. 40, pp. 4662–4676, 2024.
- [19] L. Yang et al., "Rigid-soft interactive learning for robust grasping," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1720–1727, Apr. 2020.
- [20] S. Q. Liu, L. Z. Yañez, and E. H. Adelson, "GelSight EndoFlex: A soft endoskeleton hand with continuous high-resolution tactile sensing," in *Proc. IEEE Int. Conf. Soft Robot.*, 2023, pp. 1–6.
- [21] X. Zhou and A. J. Spiers, "E-TRoll: Tactile sensing and classification via a simple robotic gripper for extended rolling manipulations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 11826–11833.
- [22] T. Yamaguchi, T. Kashiwagi, T. Arie, S. Akita, and K. Takei, "Human-like electronic skin-integrated soft robotic hand," *Adv. Intell. Syst.*, vol. 1, no. 2, 2019, Art. no. 1900018.
- [23] Z. Lin et al., "Recent advances in perceptive intelligence for soft robotics," *Adv. Intell. Syst.*, vol. 5, no. 5, 2023, Art. no. 2200329.
- [24] B. Wang, W. Guo, S. Feng, Y. Hongdong, F. Wan, and C. Song, "Volumetrically enhanced soft actuator with proprioceptive sensing," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 5284–5291, Jul. 2021.
- [25] T. Wu et al., "Vision-based tactile intelligence with soft robotic metamaterial," *Materials Des.*, vol. 238, 2024, Art. no. 112629.
- [26] O. Faris et al., "Proprioception and exteroception of a soft robotic finger using neuromorphic vision-based sensing," *Soft Robot.*, vol. 10, no. 3, pp. 467–481, 2023.
- [27] F. Wan, H. Wang, J. Wu, Y. Liu, S. Ge, and C. Song, "A reconfigurable design for omni-adaptive grasp learning," *IEEE Robot. Automat. Lett.*, vol. 5, no. 3, pp. 4210–4217, Jul. 2020.
- [28] L. Yang, X. Han, W. Guo, F. Wan, J. Pan, and C. Song, "Learning-based optoelectronically innervated tactile finger for rigid-soft interactive grasping," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 3817–3824, Apr. 2021.
- [29] D. P. Kingma, S. Mohamed, D. Jimenez Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.
- [30] X. Liu, X. Han, N. Guo, F. Wan, and C. Song, "Bio-inspired proprioceptive touch of a soft finger with inner-finger kinesthetic perception," *Biomimetics*, vol. 8, no. 6, 2023, Art. no. 501.
- [31] D. Álvarez, M. A. Roa, and L. Moreno, "Tactile-based in-hand object pose estimation," in *Proc. Iberian Robot. Conf.*, 2017, pp. 716–728.
- [32] Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama, and K. Harada, "In-hand pose estimation using hand-mounted RGB cameras and visuo-tactile sensors," *IEEE Access*, vol. 11, pp. 17218–17232, 2023.
- [33] G. Cao, J. Jiang, C. Lu, D. F. Gomes, and S. Luo, "TouchRoller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas," *Sensors*, vol. 23, no. 5, 2023, Art. no. 2661.
- [34] S. Dikhale et al., "VisuoTactile 6D pose estimation of an in-hand object using vision and tactile sensor data," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 2148–2155, Apr. 2022.