2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Exploratory Motion Guided Tactile Learning for Shape-Consistent Robotic Insertion

Gang Yan, Jinsong He, Satoshi Funabashi, Alexander Schmitz and Shigeki Sugano

***Abstract*— Intelligent robots are expected to do manipulation tasks relying on real-time sensing feedback. Especially, tactile sensing plays a more and more important role in precise manipulation tasks. For example, a 1 mm error while inserting a USB stick, which is hard to perceive visually, will result in a failed insertion or even break the USB stick. In this paper, to estimate and compensate residual position uncertainties during robotic insertion tasks, an exploration motion is introduced to acquire environment information by tactile sensing and a state-of-the-art transformer-based neural network is proposed to estimate the error distance from long-duration tactile sensing data. Our system is trained on over 2000 insertion trials with basic geometry shaped 3D printed objects. Without any prior knowledge, we achieve an 85% insertion success rate with average 5 attempts on 4 unseen daily objects relying only on tactile feedback acquired from our proposed exploratory motion. It is noteworthy that our designed exploration motion can provide insightful information about extrinsic contact information and our proposed learning model exceeds previous baselines in extracting useful information regarding the contact interaction between the grasped object and the environment.**

I. Introduction

Intelligent robots are expected to tackle a broader range of manipulation tasks in numerous scenarios, such as warehouses or even human living environments. Among various manipulation skills, inserting objects into target positions is common.

For instance, automatic warehouse robots commonly execute tasks like the typical peg-in-hole operation and are even asked to densely insert items into boxes, as these actions enhance operational efficiency and decrease costs. Moreover, humans also engage in insertion tasks under stricter conditions, employing a variety of motions, such as shape sorter toys for babies and inserting USB sticks or keys into corresponding ports. Achieving shape-consistent insertion poses great challenges, as the hole shares the exact shape but is slightly larger, requiring more precise estimation of uncertainties. Fig 1 shows an example of shape-consistent insertion.

Generally, shape-consistent insertion task requires a vision system to estimate the insertion position and orientation between the object and hole globally [37] [38]. Given simple and basic geometries, although an accurate vision sensing can align the orientation and find the approximate position, small uncertainties up to several millimeters still remain. When a grasped object encounters the environment, humans instinctively explore environmental information by rotating with respect to the collision point [12]. Torque, force, and friction information resulting from rotational motions are valuable and insightful for estimating small position errors, while this behavior is hard to be mimicked by just using robotic vision sensing. To address the limitation of vision systems and improve the estimation accuracy, several works have tried to use force feedback which will be discussed later. Different to previous works relying on force/torque sensors, in this paper, we propose to use the incorporated tactile sensing ability to capture the related torque/force/friction information between the grasped object and the environment caused by the exploratory motion and further learn how to estimate the misalignment error. In particular, a high frequency 3-axis distributed tactile sensor, uSkin [1], is used, to capture the subtle changes introduced to the elastic sensor surface caused by the collision/rotational motion w.r.t. the collision point between the object and environment. The

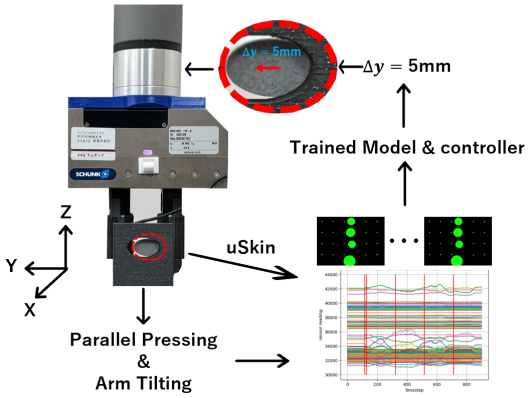

Fig. 1: **Overall system and task.** The robot is inserting a ellipse shape object into the shape-consistent hole. With the designed position misalignment (along with Y and Z axis shown in left, X axis is insertion direction), the object will collide with the edge of the hole. The robot arm will press the object parallelly and then tilt the arm, and the uSkin sensor will capture the subtle changes during the exploratory motion. The acquired tactile data, which includes the friction/torque information, is passed to the trained model and a simple controller. The model and controller will output an error estimation, which are expected to result in a successful insertion.

This research was supported by JST Moonshot R&D Grant Number JPMJMS2031, the JSPS Grant-in-Aid No. 19H01130, Sony Semiconductor Solutions Corporation, and the Research Institute for Science and Engineering, Waseda University.

The authors are with the Faculty of Science and Engineering, Dept. of Modern Mechanical Engineering, Waseda University, Tokyo 169-8555, Japan. (e-mail: yangang@fuji.waseda.jp, s-funabashi@aoni.waseda.jp, hjs898919215@akane.waseda.jp, schmitz@aoni.waseda.jp, sugano@waseda.jp)

captured sensor data stream is utilized to train our proposed state-of-art neural networks, a 3D convolution temporal-spatial transformer (Conv3D-TS-Tranformer), for estimating the position error. The 3D convolution can compress our high frequency tactile data while retaining the effective spatial-temporal features and the transformer network shares a good generalization ability among various robotic sensing modalities [30] [31]. Using the estimated position error, an iterative controller adjusts the insertion position which is expected to result in a successful insertion in subsequent attempts. During the insertion process, with the aid of the prompt sensing of uSkin, we implement a heuristic slip detection to monitor when collisions happen such that we could avoid crashing the object or hole.

In our experiment, we first collected over 2000 insertion data trials on four 3D printed basic object shapes to train and validate our proposed Conv3D-TS-Tranformer. A comparison study and ablation study were done offline using the validation dataset to show the improvement and benefits brought by our proposed model and the exploratory motion guided information. For the online experiment, we tested our system on 4 daily items with different object properties such as shape, size, surface and material. As a conclusion, our proposed exploratory guided motion could provide insightful information and our proposed model showed a better error estimation ability for long sequential time-series information. Furthermore, the online experiment on 4 unseen daily object achieves a 85% insertion success rate using on average 5 attempts, which also proves that our method is able to generalize to new objects. Finally, our dataset and code are publicly available at `https://zenodo.org/records/10706531` and at: `https://github.com/yg19930918mp/Exploratory-guided-insertion`.

Overall, our primary contributions are:

- A novel neural network model combining 3D convolution and transformer blocks is proposed to compress and extract both spatial and temporal information to estimate the position error. A comparison/ablation study shows the advantage of our proposal model.
- We provide an example of how to use an exploratory motion to acquire necessary sensing data to improve motion generation ability. The proposed exploratory motion could acquire more insightful tactile information related to torque and friction when the grasped object interacts with environment, which is barely done in other learning-based robotic tactile insertion research.
- Our dataset and code are open sourced, which is helpful for the robotic tactile sensing community.

II. Related Works

Many similar tasks, such as peg-in-hole tasks, exist to our objective, inserting an object into a shape-consistent hole. Thus, we start with an overview of solving an robotic insertion problem under uncertainties, emphasizing the necessity of introducing tactile sensors. Following this, we conduct a review of the utilization of deep learning in robotic tactile sensing, emphasizing our motivation of adopting a transformer-based approach for estimating insertion uncertainties.

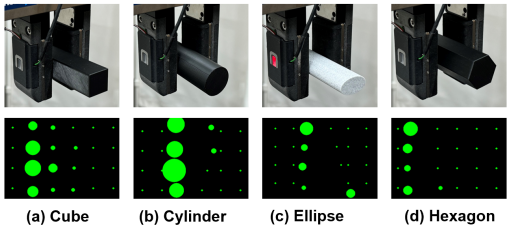

Fig. 2: **Visualization of uSkin sensor readings with our data collection objects.** Starting from the left, 3D printed cube/cylinder/ellipse/hexagon. Every green point represents a taxel within the 6x4 sensor matrix. The diameter (z-axis, normal force) and positional variation (x- and y-axis, shear force) of each green dot indicate the magnitude and direction of each 3-axis (x, y, z) measurement.

A. Robotic Object Insertion

Object insertion tasks have been a popular research topic for decades, with the peg-in-hole task being the most common format. Such tasks, i.e. contact-rich manipulation, are very challenging because the target is to generate precise position control, which can be addressed through various methods such as introducing compliance (active or passive) or employing force/torque sensing feedback.

Whitney [39] designed a RCC (remote center compliance) device to mate rigid parts based on geometric and force equilibrium conditions. Leveraging compliance, Jain et al. [2] designed a gripper employing ionic polymer metal composite (IPMC) material to illustrate the advantages of compliant fingers. Part et al. initially proposed strategies for peg-in-hole assembly without force feedback by developing compliance-based controllers and integrating different motions [4], followed by accelerated completion speed in [5]. Passive compliant joints are integrated with policy learning to emphasize the importance of passive compliance for peg-in-hole tasks in Yun S.K. [3]. Additionally, a controller capable of switching between passive compliance and active regulation is introduced in [6], surpassing fixed compliance controllers in peg-in-hole tasks. The introduction of compliance increases the generalization ability. However, the single compliance method has low efficiency since the environment information can not be leveraged to guide the adjustments of insertion.

Another class of methods learn patterns of sensor data in various insertion situations to guide the hole searching and correcting for position uncertainties. The most-used sensors are force/torque sensors. Levine et al. [7] trained a policy using images captured by an external camera on a variety of tasks, including a peg-in-hole-like assembly task. Lee et al. [8] employed deep reinforcement learning to combine image and force/torque data in peg insertion tasks, enabling generalization to unseen conditions and robustness to external perturbations. Beltran-Hernandez et al. [9] use

an off-policy, model-free reinforcement-learning method to solve peg-in-hole tasks with hole-position uncertainty. O. Azulay et al. [11] introduce haptic glance, which is similar to idea of leveraging exploratory motions to acquire more useful information, to train a curriculum learning from easier to more complex objects. However, compared to distributed tactile sensors mounted on fingertips, force/torque sensors mounted at the wrist joint can not acquire a direct contact information between the object and the fingertips, providing only a limited sensing ability.

In recent years, with the development of various types of tactile sensors, tactile sensing has been incorporated into robotic insertion tasks as it can provide more detailed and abundant sensing of the contact situation. Royo Miquel et al. [10] used a soft wrist and tactile sensors and identify the contact state transition as anomalies to search the correct insertion position. S. Dong et al. [13] [16] and S. Kim et al. [15] utilized optical tactile sensors and reinforcement learning/supervised learning to correct misalignment between objects and holes. Although the proposals effectively accomplished insertion task with unknown-shaped parts, these works used a bulky optical tactile sensor with low frequency data. However, the tactile sensor used in our research, uSkin, is a compact magnetic tactile sensor, and can provide higher frequency data.

B. Robotic tactile learning

The tactile sensor data currently available exhibits higher spatial and temporal resolutions compared to decades ago, leading to widespread use of deep learning models for processing tactile data. The learning models are used either to perceive information such as tactile related properties (texture [16] or hardness [17]) and contact state [18] [19], or to generate action directly for manipulation [20].

Considering the close relationship between the object insertion task and tactile sensing, it is reasonable to incorporate tactile sensing into this task. Here we will review the related works and then highlight the difference compared to previous learning-based tactile object insertion research. S. Dong et al. [13] estimate the misalignment between an object and the hole using CNN-LSTM [25] in a dense box packing task. Further research [16] incorporate CNN-LSTM with reinforcement learning and showed that tactile sensing enable a better generalization ability compared to force/torque sensing. S. Kim et al. [15] use the same type of tactile sensor to estimate the extrinsic contact localization and use the estimated localization to train a RL policy to search the correct hole position. Miquel et al. [10] leverage a soft wrist and tactile sensor to detect contact state transitions as anomalies to locate the optimal insertion position. R. Okumura et al. [21] introduce a world model for mapping tactile images to physical coordinates and implement the model for a connector insertion task.

Compared to previous works, our work mainly has 2 distinguishing aspects. Firstly, we introduce a transformer-based model to estimate the error and compare our proposal to the widely-used CNN-LSTM [13] [16] [23] [24]. Secondly, with the introduction of exploratory guided motion to gather additional information, each data sample includes a longer time dependency (up to 750 time- steps), making it more challenging to extract relevant useful information.

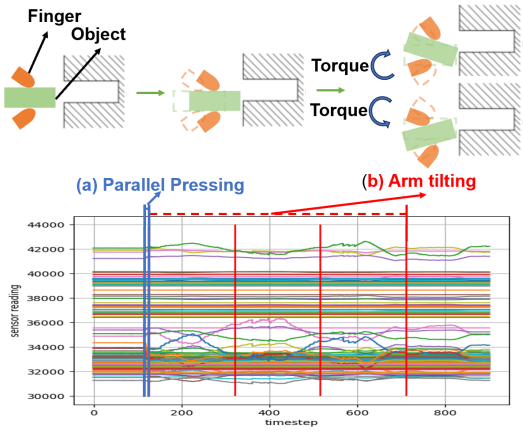

Fig. 3: **Top id exploratory motion. Bottom is corresponding tactile sensor reading.** (a) Parallel pressing motion. (b) Arm tilting motion.

III. Proposed Method

Human are able to identify object properties and estimate a correction signal through a tactile exposure in insertion processes without visual feedback especially when the misalignment is relatively small. Our tactile sensors are prompt and sensitive and are capable of detecting subtle changes when the grasped object collides with the environment or when the object is rotated relative to the collision point. Using the benefits of our tactile sensor, we propose to introduce an exploratory motion similar to human behaviors to capture the extrinsic contact information when the grasped object interacts with the environment. Further, to leverage such subtle information to estimate the error direction/magnitude, we propose to compress (embed) the spatial-temporal tactile data using 3D convolution and then use a transformer-based network to extract useful information. In next part, we will first formulate our task and then detail our proposal.

A. Problem formulation

We use a robotic arm equipped with a parallel gripper and a distributed tactile sensor patch is mounted on each fingertip, as shown in Fig. 1. The system will attempt to insert basic geometry shaped object into a shape-consistent hole with uncertainties while the hole has a relatively small clearance (1.5 to 2mm). For the action space of grasped objects, assuming that the orientation uncertainty is more related to a visual sensing problem, we only considered the position uncertainties (Y and Z axis shown in Fig. 1). During the whole insertion, the system could only acquire tactile measurements. The insertion goal is, therefore, to align the object and insert it into the hole without any prior knowledge of the object shape nor the environment by just leveraging the tactile measurements and the learning model.

The described task above represents a typical regression task. Consequently, our proposed network retains a regression output format, with the output consisting of continuous values that represent the estimated error distance in Y and Z direction.

B. Exploratory Motion

Humans are used to obtain external environment properties via purposeful manual exploration relying on haptic feedback. Although the human haptic system includes muscles, tendons, joints and skin, in our research we mainly focus on skin-like tactile sensors as it is accepted that the cutaneous feedback is the most prompt and abundant information source [12]. Taking inspiration from human motion during insertion, we propose to add the exploratory motion after the object collides with the edge of the hole.

Our proposed motion includes two parts, **parallel pressing** and **arm tilting**, as shown in Fig. 3.

Parallel pressing motion (Fig. 3(a)). As the baseline exploration motion which is used in other tactile insertion learning research [13], our robot arm would push the grasped object towards the target position until the object collided with the edge of the hole and the tactile sensor readings reached a shear force threshold. A successful insertion where the object could go through the hole without any collision could be also identified by monitoring tactile readings. While inserting, due to the sensor surface compliance and different intersection situation between the object and hole, a slightly torque could happen accordingly and our distributed sensor point would capture this informative signals.

Arm tilting motion (Fig 3(b)). Inspired by human behaviors [12], object tilting is deliberately exerted by the arm in order to sense the edge of the hole. We believe the tilting motion could provide a more obvious information such as the friction and torque between the object and surface surrounding the hole. Also, our prompt and high frequency distributed tactile sensor is able to capture this feature. Hence, the arm is tilted back and forth to two directions with respect to the surface normal. The center of each tilt is approximately the contact point with the hole. While the large tilting motion may result in a unexpected posture variation of the object, we select a 10 degree tilting in each direction and we observed that the resulting posture variation could be compensated by the tactile sensor surface compliance and the grasping strength.

C. Learning Model

Using exploratory motions, we capture a time-series of distributed tactile data along with an added error distance $(\Delta Y, \Delta Z)$ with respect to the known correct hole center position. Our aim is to compress the data and extract valuable information to estimate the error distance.

As shown in Fig. 2, the uSkin sensor patch used in our experiment consists of 24 sensing taxels, each taxel providing 3-axis readings with a 100 Hz frequency. Despite not reaching the spatial resolution of RGB tactile images obtained

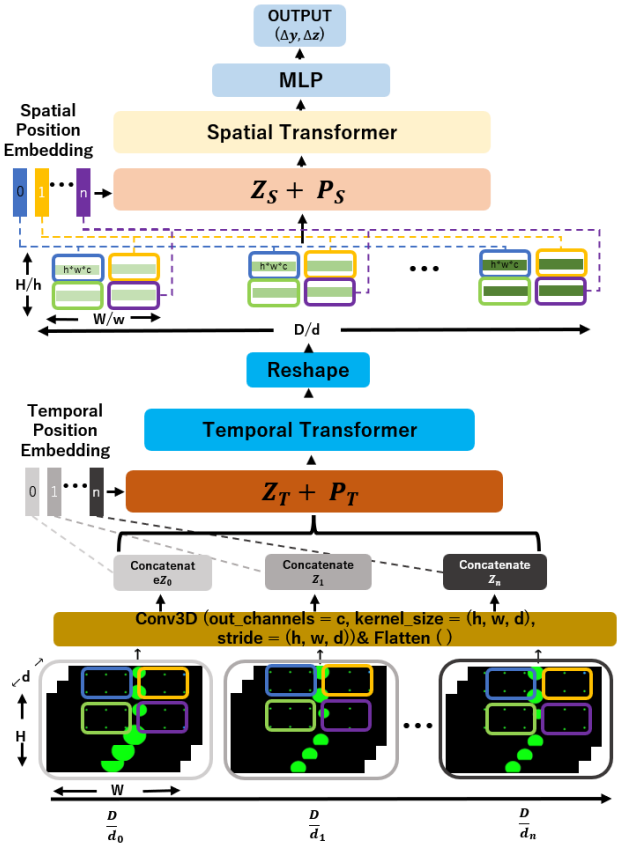

Fig. 4: **Our proposal model.** The spatial-temporal input data is first compressed by 3D convolution on height, width and depth. The compressed tokens are passed through a temporal transformer and spatial transformer while the temporal position encoding (large grey rectangles in the bottom) and spatial position encoding (colorful small rectangles in the bottom) are added respectively. The processed features are connected to outputs values, ΔY and ΔZ, by a multi-layer-perceptron.

from optical tactile sensors, our data retains complex spatial-temporal dependency information. Therefore, the design of our network should consider how to extract spatial-temporal features. In previous works, the CNN-LSTM [25] [26] [27] was widely used to extract both spatial-temporal information for robotic tactile learning. Our previous research [28] [27] validated that despite being trained from scratch, our CNN-LSTM achieves satisfactory classification performance. Furthermore, our proposed self-attention mechanism surpassed the CNN-LSTM in handling long sequential data of over 100 time-steps. However, in our current experiment the data sample has a longer sequence of up to 750 time-steps. We found that the CNN-LSTM showed worse performance. Thus we propose to introduce the self-attention mechanism into our task.

To leverage the power of self-attention mechanism, a straight-forward idea is to treat our tactile data as a fixed length video and each frame consists of 2 (sensor patches)

* 6*4 (sensing points) * 3 (axis) = 144 data. By embedding each frame into vectors (called tokenization in the rest of the paper), it becomes naturally consistent with typical transformer input structure. However, this idea has several limitations. Firstly, tokenization is generally accomplished through convolution and fully-connected layers. Nevertheless, embedding long sequential data like ours (up to 750 time-steps) using this approach would introduce a significant number of additional neural network parameters, which would heighten training difficulty and diminish inference speed. Additionally, tokenization using CNN and FNN to compress and flatten each individual frame as input for the transformer might maintain spatial features within each frame, but it might disturb the spatial-temporal relationships between different frames. For example, variations in shear force readings on the same taxel position during successive time-steps could convey important information, such as the perception of contact friction or torque, which would be ignored if we compressed and flattened single frames separately. Lastly, the self-attention mechanism requires substantial computation of multi-dimensional linear algebraic arrays. As the length of input data increases, it dramatically consumes GPU memory (up to 25 GB GPU memory space when using PyTorch), constraining parameter tuning options like batch size selection and placing a heavier burden on the GPU.

Facing the above challenges, we propose a novel Conv3D-Temporal-Spatial transformer accordingly, based on related works. Given the issues outlined above, the proposed model is tasked with initially compressing the lengthy sequential data while retaining effective features, and subsequently extracting the preserved spatial-temporal dependencies to estimate the position error values. The whole model is shown in Fig. 4.

3D Convolution Embedding Previous research on Vision transformer (ViT) [32] involved the tokenization of a 2D image into a collection of patch tokens. This method entails dividing the image into a sequence of uniform-sized, non-overlapping patches with spatial extent, which are subsequently flattened into 1D tokens and linearly projected into d dimensions. This tokenization process could be treated as a 2D convolution with a kernel size of (h × w) and strides (h, w) across the respective dimensions. As mentioned before, our data consists of a series of distributed tactile data with T (750) time steps, H * W (6 * 4) sensing points and C (3) channels. Inspired by ViT and our previous work, we propose to tokenize our recorded data $X \in R^{T*H*W*C}$ by extending the 2D convolution approach to 3D convolution, where a 3D kernel with size (d×h×w) is applied at stride (d, h, w) across temporal and spatial dimensions. Consequently, $N = [\frac{D}{d}, \frac{H}{h}, \frac{W}{w}]$ non-overlapping tokens are extracted with size $X \in R^{d*h*w*c}$ in this manner and then flattened. By setting $d > 1$, all extracted tokens contain spatial-temporal information. As the computation cost of self-attention layers increases quadratically concerning the length of the token sequence, selecting suitable values for t, h, and w are important. We will detail our selection in the later experiments part.

In the end, our tokenization strategy preserves both spatial and temporal information while efficiently compressing the data length, thus minimizing unnecessary computation costs.

Temporal Encoder The extracted patches descried above $Z_T \in R^{N_D*N_H N_W N_C}$ are used to construct model inputs as follows

$$Z_T^{input} = Z_T + P_T \in R^{N_D*N_H N_W N_C}$$

where the $P_T \in R^{N_D*N_H N_W N_C}$ is a set of learned positional encoding vectors and are added to all $N_H N_W N_C$, to encode the absolute temporal information. The absolute temporal position is represented by the different level of grey rectangles shown in the bottom of Fig. 4. Within the typical transformer, there are L blocks consisting of alternating layers of Multiheaded Self-Attention (MSA) and residual Multi-Layer Perceptron (MLP). Prior to each layer, inputs are normalized with layer-norm [33]. The MLP block includes two layers of linear projection followed by GELU non-linear activations [34]. Unlike CNN architectures, where spatial dimensions shrink and channel numbers increases, transformers maintain isotropy, with all feature maps retaining the same shape throughout the network. Empirically, we selected L = 1 in our later experiments.

Spatial Encoder We now divide the second dimensions of temporal encoder output into position patches and channels, obtaining a list of patch features $Z_s \in R^{N_D N_C * N_H N_W}$. The positional embedding is given according to the divided patch positions (colorful rectangles shown in the bottom of Fig. 4). In a similar spirit, the input to the spatial encoder becomes:

$$Z_S^{input} = Z_S + P_S \in R^{N_D N_C * N_H N_W}$$

where $P_S \in N_D N_C * N_H N_W$ are respectively added to all spatial representations. With all feature maps retaining the same shape throughout the network, another two layer MLP is followed to project the features into the estimated error distance.

A schematic representation of the our proposed architecture can be seen in Fig. 4. Our proposal is compared against several baseline, which will be detailed in later sections.

IV. Experiment and Results

A. System Overview

We build our experimental setup to automatically collect the training and validation data as shown in Fig. 1. The setup consists of the two-finger parallel gripper WSG-110 and a 4*6 uSkin sensor patch is mounted on each fingertip. The internal force sensing function of WSG-110 is not reliable because we are using customized fingers during the whole experiment, so the grasping strength and detection of slip is done by monitoring uSkin readings. The gripper is mounted on a UR3e robot arm. During the experiment, real-time data stream of uSkin is recorded with 100Hz once the object collides with the edge of the hole.

B. Data Collection

Objects and Holes As shown in Fig. 2, four 3d printed objects with basic geometry shapes are selected to collect training and validation data. The holes w.r.t each object share

TABLE I: Motion/Model Comparison (unit: mm^2)

Model	CNN-LSTM	CNN-LSTM-AE	Conv3D-TS
Validation Loss (parallel pressing motion)	7.394	6.793	5.998
Validation Loss (proposed motion)	6.099	5.966	4.1

the same shape and the hole clearance is 1.5∼2mm larger than the size of the object. We treat this clearance as a challenging clearance since even when the robot is controlled by a human operator, it takes several attempts to insert the object successfully given this clearance.

Uncertainties Since the hole is designed 1.5∼2mm larger than the object size, the positional uncertainties of insertion are defined as 3∼10mm distanced from the center of the hole in 36 directions (every 10 degree).

Motion and Size of Dataset During our automatically data collection process, each object is inserted with a random combination of distance and direction error about 500 times, resulting in a dataset including about 2000 times insertions. For each insertion attempt, tactile data is recorded from the start of insertion including the parallel pressing motion and the arm tilting motion and the continuous 750 time-steps from the happening of collision to the finishing of the arm tilting are sampled as one training/validation sample. Before each insertion, baseline data is recorded before the start of insertion and the baseline data is used to pre-process the raw data. The insertion speed is fixed to avoid unnecessary variation.

C. Model Training

Data Preprocessing Our input data is scaled to [-1, 1] by using the empirical sensor reading range in our experiments and the baseline data recorded before the start of insertion.

Training setting We use the ADAM optimizer, MSE loss function, NVIDIA 3090 GPU and Pytorch package for offline experiments. 85%/15% of our dataset is used for training/validation. For network details, please refer to our code at the publicly available github.

D. Offline Studies

For the offline studies, we use the mean square error (MSE) as the evaluation metric. Achieving a lower MSE loss indicates that the estimation moves closer to the ground truth.

Model Comparison To perform a model comparison study, two baseline models are selected. The first baseline model is CNN-LSTM, which is widely used in previous research to extract both spatial and temporal features. Following our previous work, since our data from our tactile sensor is not an RGB image, we do not leverage the backbone models pretrained on an RGB image dataset. Instead, all our models are trained from scratch. For the second baseline, we add an auto-encoder structure into the CNN-LSTM to augment the feature extraction ability. An autoencoder is a type of artificial neural network used for unsupervised learning. It is primarily designed to learn efficient representations of data, typically by encoding the input into a lower-dimensional latent space and then decoding it back to the original input space. It can learn meaningful and powerful features or representations of the input data, which can be useful for robotic manipulation tasks [35] [36]. In particular, a de-convolution decoder and de-LSTM decoder is added separately after the CNN and LSTM part to reconstruct the corresponding inputs. Both of them are trained using MSE loss together from scratch with the CNN-LSTM.

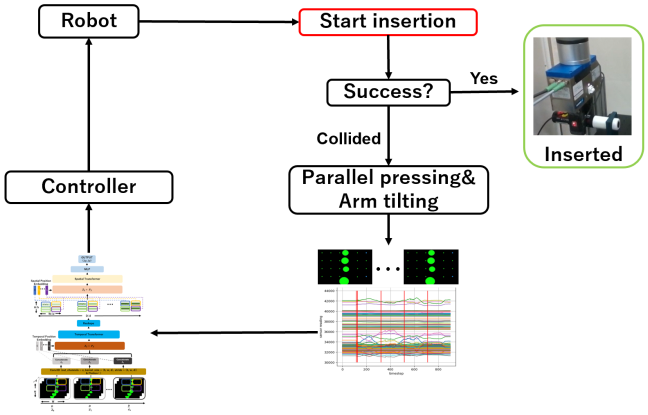

Fig. 5: **Work flow for real-time experiments.** The robot initiates the insertion process, and if successful, it proceeds to place the object back to the reset device. In the event of a collision between the object and the environment, parallel pressing and arm tilting motions are activated, and tactile data are captured and processed by sensors. The processed sequences are then forwarded to our proposed neural networks. The estimated error is passed to a simple controller, which determines the adjustment of the insertion position. If insertion at the new position is successful, the process is done. Otherwise, a new error is detected again by the sensors. This loop continues until success is achieved or the maximum attempts threshold (20) is reached.

The validation losses are shown in Table I. Here we show that the estimation loss of our proposed Conv3D-TS transformeris is 4.1 where the absolute error is estimated within $-1.4mm \sim 1.4mm$ on each axis. However, the baseline, CNN-LSTM, could only reach estimation loss 5.966 with auto encoders and 6.099 without auto encoders. Compared to baseline, our proposed model could improve the estimation error up to 33%.

Motion Comparison In addition to our proposed network model, we also believe the friction/torque information happening during exploratory motions (parallel pressing motion and arm tilting motion) could be captured by our tactile sensor and this information could improve the estimation accuracy. In previous research such as [15], only the pressing motion is used to acquire environment information. Therefore, we treat the parallel pressing motion as baseline motion and compared our proposal exploratory motion with the baseline motion. To show the effectiveness of our proposed exploratory motion, we do a motion comparison study on CNN-LSTM, CNN-LSTM-AE and our proposed model, Conv3D-TS transformer.

TABLE II: Ablation study (unit: mm^2)

Model	Conv3D	Conv3D-temporal transformer	Conv3D-spatial transformer	Proposal
Validation Loss	6.014	4.246	4.184	4.1

The motion comparison result is shown in Table I. For all models the performance dropped when only using the parallel pressing motion. It is evident that the extra contact information brought by our proposed motion is beneficial for the estimation of misalignment.

Influence of compression As discussed before, we believe that a suitable compression of our long sequential spatial data by using 3D convolution is able to keep the effective features for error estimation. Therefore the tuning of parameters of 3D convolution kernel, height (h), width (w) and t (depth), are important where h/w control the receptive field of spatial aspect and t control the receptive field of temporal aspect. To show how the size of h and w and t influence test performance, we do a comparison study by fixing all other parameters while only changing h/w or t based on our proposed model.

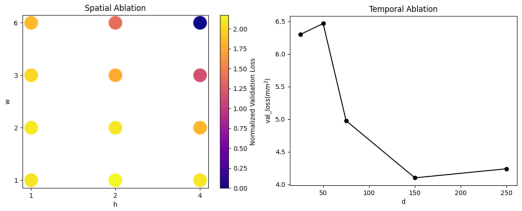

Fig. 6: **Ablation study of 3D convolution receptive field.** Left: Study about height and width. X-axis is height and y-axis is width. The color of point represents the estimation loss. Right: Study about depth. X-axis is depth and y-axis is the estimation loss.

The corresponding results are shown in Fig. 6. The left of Fig. 6 shows how the height and width influence the learning result where the depth is fixed to 150. Since our design motivation is to divide inputs into non-overlapping patches, the height and width could be only set to [1,2,4] and [1,2,3,6] separately. The cooler color represents a smaller estimation error. It is obviously that as the receptive area (h*w) increases, the estimation error decreases, which proves that the larger receptive area is beneficial for our task. Similarly,in the right of Fig. 6 we found that as the convolution depth increases, the estimation loss decreases. However the optimal depth is 150 with the estimation loss 4.1.

Ablation Study Although our proposed model outperforms the baseline models, it is important to know why our model could bring such improvement. Thus we do an ablation study. As mentioned before, our model includes a compression part, temporal transformer and spatial transformer. In our ablation study, we compare the validation performance when we remove either temporal transformer/spatial transformer or both. The rest of the parameters are fixed.

The ablation study results are shown in Table II. Without either temporal transformer or spatial transformer, the validation loss slightly dropped 0.084 and 0.146 respectively. However, without both, the pure 3D convolution model could only reach 6.014 validation loss, which shares a similar performance to CNN-LSTM and proves that the transformer blocks are quite necessary to extract the spatial and temporal relevance.

TABLE III: Real robot experiment results

object name	White pen	Red pen	Glass bottle	Toothpick	
Object picture					Total
Successful rate	5/5	3/5	5/5	4/5	17/20
Average attempts for success trials	3.2	8.7	2.2	8.0	5

E. Online Robot Experiment

Finally, to test our proposed system and evaluate whether it is able to generalize to unseen daily objects, we do an online real-time insertion experiment. Based on the offline comparison results, our proposed model is selected to do the online experiment since it outperforms other baseline models evaluated by MSE loss. Four daily objects with basic geometries but various properties are selected as shown in Table III. For example, the red mark pen has a non-fixed diameter and different texture, and the toothpick box is deformable. The glass bottle and mark pen are heavier than the 3D printed objects used for training. All of them have a different size compared with the training objects.

Instead of collecting one-time insertion error data (as we did for training/validation data collection), the robot arm will keep repeating attempting inserting until successfully insertion or the number of attempts exceeds 20, see Fig. 5.

Control Strategy Since our training data are not collected for inserting objects with continuous attempts, it is necessary to use a strategy to regulate the real-time experiments. After some trial and error, two regulations were implemented. Firstly, we clip the prediction values by multiplying them with 0.6. We found that without the clipping, the insertion position will be dragged to be distanced from the hole once an unreliable prediction happened and thus increasing the difficulty of later attempts. Secondly, to lower the frequency of unreliable predictions, we combine multiple checkpoints of the trained model by averaging the output of multiple checkpoints.

Online experiment result The online experiment results are shown in Table III. Each object is inserted 5 times and we reached a 85% successful insertion rate with 5 modifications on average. The white mark pen and glass bottle reached a 100% success rate as they have a regular shape and hard, smooth surface. However, the insertion of varied diameter red pen and the deformable toothpick box failed several times. We believe the unseen properties bring some difficulties for estimating the error distance. Even for successful trials, these two objects require more attempts. Overall, our method is able to generalize to unseen objects while somewhat struggling with soft and irregular diameter shaped objects.

V. Conclusion and Discussion

In this paper, we proposed to introduce an exploratory motion to gather more insightful tactile information and leverage this spatial-temporal information by using a Conv3D temporal-spatial transformer. The proposed neural network model merges 3D convolution and transformer blocks to compress and extract spatial and temporal information. This integration enables the estimation of position error more accurate than previous baselines. Based on the estimated error and a simple controller, our system is able to insert new objects into a shape-consistent hole within 5 attempts on average to overcome a relatively small misalignment.

Here we also list the limitations of our proposed method and possible future solutions. Our method does not take into account the sequence of past actions and observations. This will lead to a accumulated error in some situations and a waste of previous observations. We believe that by introducing reinforcement learning, we would be able to solve this problem. The objects used in our experiments all have a flat base and most of them are rigid. It is necessary to investigate the generalization ability given a more complex object set. Further, the action space is limited into Y and Z translation. A better inserting strategy should be able to overcome both position and posture misalignment. We believe this could be solved by using other sensing modalities such as vision and compliant hardware in the future.

References

[1] Tomo, T.P. et al., "A New Silicone Structure for uSkin - A Soft, Distributed, Digital 3-Axis Skin Sensor and Its Integration on the Humanoid Robot iCub," *IEEE Robotics and Automation Letters,* vol. 3, pp. 2584-2591, July 2018.

[2] R.K. Jain et al., "SCARA based peg-in-hole assembly using compliant IPMC micro gripper," *Robotics and Autonomous Systems,* Volume 61, Issue 3, 2013, Pages 297-311.

[3] Yun S.K. et al., "Compliant manipulation for peg-in-hole: Is passive compliance a key to learn contact motion?"(2008), *IEEE International Conference on Robotics and Automation,* art. no. 4543437, pp. 1647 - 1652.

[4] Park H. et al., "Compliance-Based Robotic Peg-in-Hole Assembly Strategy Without Force Feedback", (2017), *IEEE Transactions on Industrial Electronics,* 64 (8), art. no. 7914743, pp. 6299 - 6309.

[5] Park H. et al., "Compliant Peg-in-Hole Assembly Using Partial Spiral Force Trajectory with Tilted Peg Posture", (2020), *IEEE Robotics and Automation Letters,* 5 (3), art. no. 9109673, pp. 4447 - 4454.

[6] Ren T. et al., "Learning-based variable compliance control for robotic assembly", (2018), *Journal of Mechanisms and Robotics,* 10 (6), art. no. 061bibitem008 Journal of Machine Learning Research,, vol.17, no.1, pp 1334-1373, 2016..

[7] Sergey Levine et al., "End-to-End Training of Deep Visuomotor Policies", *The Journal of Machine Learning Research,*, vol.17, no.1, pp 1334-1373, 2016.

[8] M. A. Lee et al., "Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks," in *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 582-596, June 2020.

[9] Beltran-Hernandez et al. 2020. "Variable Compliance Control for Robotic Peg-in-Hole Assembly: A Deep-Reinforcement-Learning Approach" *Applied Sciences* 10, no. 19: 6923.

[10] Royo Miquel et al., "Learning Robotic Assembly by Leveraging Physical Softness and Tactile Sensing", (2023), The IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).

[11] O. Azulay et al., "Haptic-based and SE (3)-aware object insertion using compliant hands," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 208–215, 2022.

[12] R. L. Klatzky et al., "Identifying objects from a haptic glance," *Percep. Psychophys.*, vol. 57, pp. 1111–1123, 1995.

[13] S. Dong et al., "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 6437–6443.

[14] S. Kim and A. Rodriguez, "Active extrinsic contact sensing: Application to general peg-in-hole insertion," in *IEEE International Conference on Robotics and Automation*, 2022.

[15] S. Dong et al., "Tactile-Based Insertion for Dense Box-Packing," *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

[16] A. Geier et al., "Deep GRU-ensembles for active tactile texture recognition with soft, distributed skin sensors in dynamic contact scenarios," *2020 IEEE/SICE International Symposium on System Integration (SII)*.

[17] W. Yuan et al., "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor," *2017 IEEE International Conference on Robotics and Automation (ICRA)*.

[18] G. Yan et al., "Detection of Slip from Vision and Touch," 2022 *International Conference on Robotics and Automation (ICRA)*.

[19] G. Yan. et al,"SCT-CNN: A Spatio-Channel-Temporal Attention CNN for Grasp Stability Prediction", 2021 *IEEE International Conference on Robotics and Automation (ICRA),*.

[20] R. Calandra et al., "More Than a Feeling: Learning to Grasp and Regrasp Using Vision and Touch," in *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300-3307, Oct. 2018.

[21] R. Okumura et al., "Tactile-Sensitive NewtonianVAE for High-Accuracy Industrial Connector Insertion," 2022 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 2022, pp. 4625-4631

[22] Vaswani, A. et al., "Attention is all you need", (2017) *Advances in Neural Information Processing Systems*.

[23] Yuan, W. et al., "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor", *IEEE International Conference on Robotics and Automation*, 2017.

[24] Yuan, W. et al., "Active clothing material perception using tactile sensing and deep learning", (2018) *IEEE International Conference on Robotics and Automation*, pp. 4842-4849.

[25] Tara N. Sainath et al., (2015) "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks", *IEEE International Conference on Acoustics, Speech and Signal Processing*.

[26] K. He et al., "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778.

[27] Hochreiter, S. et al., (1997), "Long short-term memory", *Neural computation*, vol. 9, no. 8, pp. 1735–1780.

[28] Gang Yan et al., (2021). "SCT-CNN: A Spatio-Channel-Temporal Attention CNN for Grasp Stability Prediction", *IEEE International Conference on Robotics and Automation*.

[29] Gang, Y. et al. "Detection of Slip from Vision and Touch", *IEEE International Conference on Robotics and Automation*, 2022.

[30] Anthony Brohan et al., "RT-1: Robotics Transformer for Real-World Control at Scale", 2023, 2212.06817, arXiv, cs.RO.

[31] Anthony Brohan et al., "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control", 2023, 2307.15818, arXiv, cs.RO.

[32] Alexey Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", 2021, 2010.11929, arXiv, cs.CV

[33] Jimmy Ba et al., "Layer normalization", ArXiv, abs/1607.06450, 2016

[34] Dan Hendrycks et al., "Gaussian error linear units (gelus)". arXiv: Learning, 2016.

[35] Chelsea Finn et al. 2016. "Deep spatial autoencoders for visuomotor learning". In *2016 IEEE International Conference on Robotics and Automation (ICRA)*.

[36] H. van Hoof et al., "Stable reinforcement learning with autoencoders for tactile and visual data," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* .

[37] B. H. Yoshimi et al., "Active, uncalibrated visual servoing," Proceedings of the *1994 IEEE International Conference on Robotics and Automation*.

[38] G. Morel et al., "Impedance based combination of visual and force control," Proceedings. 1998 *IEEE International Conference on Robotics and Automation* (Cat. No.98CH36146).

[39] D.E. Whitney, "Quasi-static assembly of compliantly supported rigid parts", *ASME J. Dynamic Systems, Measurement and Control*, 1982.