IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Modality Attention for Prediction-Based Robot Motion Generation: Improving Interpretability and Robustness of Using Multi-Modality

Hideyuki Ichiwara[1,2], Hiroshi Ito[1,2], Kenjiro Yamamoto[1], Hiroki Mori[3] and Tetsuya Ogata[2,4]

***Abstract*—We developed a modality attention motion generation model on the basis of multi-modality prediction. This model provides interpretability about modality usage and demonstrates robustness against disturbances. We used a hierarchical model consisting of low-level recurrent neural networks (RNNs) for processing each modality individually and a high-level RNN that integrates the multi-modality. This integration is achieved by efficiently gating multi-modality and inputting it to the high-level RNN. We verified the interpretability and robustness of the task of inserting a furniture part, which consists of the "approach" phase to bring the wooden dowel closer to the hole and the "insertion" phase. While the proposed model achieves the same task success rate as the conventional model, it clarifies that it refers to vision during "approach" and force during "insertion," providing interpretability regarding modality use. Furthermore, in contrast to the non-modality attention model, whose task success rate drops significantly under disturbance, the proposed model enhances robustness against disturbances to modalities it does not direct attention during the task, resulting in a consistently high success rate ($\simeq 90\%$).**

***Index Terms*—Deep Learning in Grasping and Manipulation, Neurorobotics, Learning from Experience**

I. Introduction

The human brain skillfully processes sensory signals from various sensory organs in the body, such as eyes, ears, skin, and muscle spindles [1]. These sensory processes enable humans to perform a variety of tasks. Methods have also been proposed for robots that use multi-modality to make robots perform tasks in the same way humans do. For example, methods using reinforcement learning (RL) [2][3][4], and those using supervised learning [5][6][7] have been proposed. For methods using RL, trial and error is required and sample efficiency is a problem, but many methods to improve it have been proposed [8][9][10]. However, trial and error is still necessary, so there is still a risk of damaging the work object or yourself in tasks that come into contact with objects. Furthermore, it requires time-consuming reward design and tuning of various hyperparameters for each task [11], and adjusting them requires trial and error. On the other hand, supervised learning-based deep predictive learning (DPL) has been proposed [5][12][13]. This method uses human demonstration data to train a model for the purpose of multi-modality time-series predicting. Then, real-time sensor values are input to the model and predicted command values are input to the robot to generate motion. Therefore, trial and error that leads to reward design or damage risk is eliminated, and the risk is relatively low.

Motion generation using multi-modality uses a variety of sensors. However, it is not easy to combine modalities with different physical meanings, time constants, and amounts of information to learn robot motion. There is also the risk of overfitting for a single modality. For a single modality, there have been several discussions about methods to improve robustness against disturbances and sensor values that differ from those during learning. In particular, there are several studies targeting visual information, which has a high amount of information and a high degree of abstraction [2][14][15]. Among them, several studies have improved robustness against background changes and sample efficiency through efficient visual processing that introduces the concept of attention in vision [14][16].

Attention is not only for vision but also for multi-modality. Depending on the situation, it is not necessary to pay attention to all modalities at all times, and the modality of focus is switched [17]. We consider that introducing modality attention to robots is important for flexible and robust motion. It also contributes to interpretability, which is important in terms of industrial application, by providing humans with information about the referred modality. In this study, we propose a model of real-time motion generation of robots that introduces the concept of modality attention and describe an evaluation experiment using a real robot for the task of approaching and inserting a chair part. This method uses a hierarchical model that uses lower recurrent neural networks (RNNs) that process each modality individually and a higher RNN that integrates the multi-modality. Furthermore, by weighting and inputting each modality input to the upper neural network (NN), the modality used for motion generation can be self-adjusted. It is possible to interpret how much each modality is referenced by checking the weights. Our main contributions are as follows:1) Development of a real-time robot motion generation model that introduces the concept of modality attention. 2) Evaluation of the interpretability of modality use and robustness against

[1]Hideyuki Ichiwara, Hiroshi Ito and Kenjiro Yamamoto are with Research & Development Group, Hitachi, Ltd., Ibaraki, 312-0034, Japan {hideyuki.ichiwara.bn, hiroshi.ito.ws, kenjiro.yamamoto.bq}@hitachi.com

[2]Hideyuki Ichiwara, Hiroshi Ito and Tetsuya Ogata are with the Graduate School of Fundamental Science and Engineering, Waseda University, Tokyo 169-8555, Japan

[3]Hiroki Mori is with the Future Robotics Organization, Waseda University, Tokyo 169-8555, Japan mori@idr.ias.sci.waseda.ac.jp

[4]Tetsuya Ogata is with the Waseda Research Institute for Science and Engineering (WISE), Waseda University, Tokyo 169-8555, Japan ogata@waseda.jp

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

disturbances of the proposed model for the task of inserting a chair part. 3) Applicability to other tasks was shown through demonstrations of the outlet plug insertion and marker wiping tasks.

II. Related Work

A. Autonomous robot manipulation using multi-modality

A wide range of tasks, such as complex tasks consisting of multiple sequences [13][18], soft object manipulation [19][20], and contact-rich object manipulation [6][12][21][22][23], have been achieved using multi-modality. Furthermore, to deal with differences in the physical meaning and time-varying characteristics of each modality, [24] proposed a hierarchical motion generation model using two types of RNN, and performed smooth motion generation using jerk as an index. In these studies, it is considered that task execution by robots was achieved by appropriate integrated processing of multi-modality. However, although it can be inferred from the results of task execution that multi-modality was effectively used, it was not clear which modality was used and to what extent.

In addition, there are several studies [25][26] that use multi-modality for the peg insertion task in this study for quantitative evaluation. In [25], the policy is acquired through RL. For its representation learning, it was proposed to predict optical flow, the presence or absence of contact, and whether two sensor streams are temporally aligned. The policy obtained on the basis of the method was to search for holes while touching the target. Due to space constraints and damage risk, there are few practical scenarios where this is acceptable. Depending on the reward design, different policies may be obtained, but the method that predicts the three pieces of information may function effectively assuming the previous policy is obtained. On the other hand, in supervised learning-based frameworks such as DPL, it is possible to reduce the risk of damage by using demonstration data to avoid searching while touching during approach. In the framework, focusing on the unreliability of raw sensor values, [26] proposed multi-step command prediction and feature extraction from force sensor data. Furthermore, it was shown to be robust against changes in the target's position. However, the evaluation was limited to a nominal environment, and the robustness against sensor disturbance is unknown.

B. Attention for robot manipulation

There are several studies about attention for vision [14][16][27]. For vision processing, autoencoder was used to map vision to low-dimensional feature values, which were used to generate motion [13][22][28]. On the other hand, [2][14][16] proposed a framework for extracting positions in an image as a feature value and improved visual interpretability. [12][14] showed both interpretability for vision and robustness against background changes, etc., and also demonstrated applicability to the manipulation of flexible objects accompanied by dynamic visual changes.

Regarding attention for multi-modality, there are several studies that introduce the concept of attention to the integration of vision and tactility [29][30]. [29] integrates them on a self-attention basis in a model that predicts the success/failure of grasping tasks. Anzai et al. [30] developed modality attention that can interpret which modality is referred to and to what degree in the in-hand pose estimation of an object. In [30], tactility and images are used as modalities, the weight of each modal is obtained from the feature value, and the weighted feature value is used for pose estimation. It can be said that it is bottom-up attention that depends on the sensor value. They also achieved noise robustness by adding noise to each modality during training. This takes advantage of the fact that the task of pose estimation can be accomplished to some extent with either tactility or images. There is also the idea [31][32] of not referring to untrusted modalities when they are present, although this does not quite match the idea of attention. They assume that the task under consideration is feasible with few (or single) modalities in a redundant sensing environment. On the other hand, there have been few discussions about modality attention in robot manipulation, where the modalities required in a task are expected to change depending on the time and task state. For example, even when simply grasping an object, vision and tactility are used in approaching and grasping, respectively. The purpose of this research is to propose a modality attention method in robot manipulation and to show that the method is effective for real robot task execution.

III. Method

In the proposed method, DPL is used to reduce the risk of damage to the work object and the robot. DPL uses NNs to predict multi-modality at the next time $t+1$ that includes control commands of the robot from those at the current time t. The predicted commands are sent sequentially to the robot to generate the motion. The model is trained using the time-series demonstration data. Fig. 1(A) shows the overall structure of the model. The inputs are an image $\boldsymbol{v}_t$ at time t, control commands to the robot $\boldsymbol{c}_t$, and other modalities $\boldsymbol{m}_t$. The outputs are each predicted modality $\hat{\boldsymbol{v}}_{t+1}, \hat{\boldsymbol{c}}_{t+1}, \hat{\boldsymbol{m}}_{t+1}$ at time $t+1$. Note that the hat indicates the output of the NN, and the absence of the hat indicates the true value (measured value). The details of the loss function are shown in Section III-C, and the model updates the weights by backpropagating the modality prediction errors.

A. Visual attention

Human visual attention consists of bottom-up and top-down, and they function while interacting with each other [33]. Bottom-up attention is passive, directed toward salient stimuli, such as those with significant contrast differences. Top-down attention is active, with prior knowledge of the stimulus to choose from and biased attention to a specific stimulus. Fig. 1(A) shows the visual attention model [14][24] used in this study. This model was inspired by this idea and builds bottom-up attention from multiple convolutional NNs (CNNs) and soft argmax [34] that obtains the coordinates of the most intense pixels from the feature map. Top-down attention is built from image predictions. Both attentions interact through learning and inferencing through the RNN.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

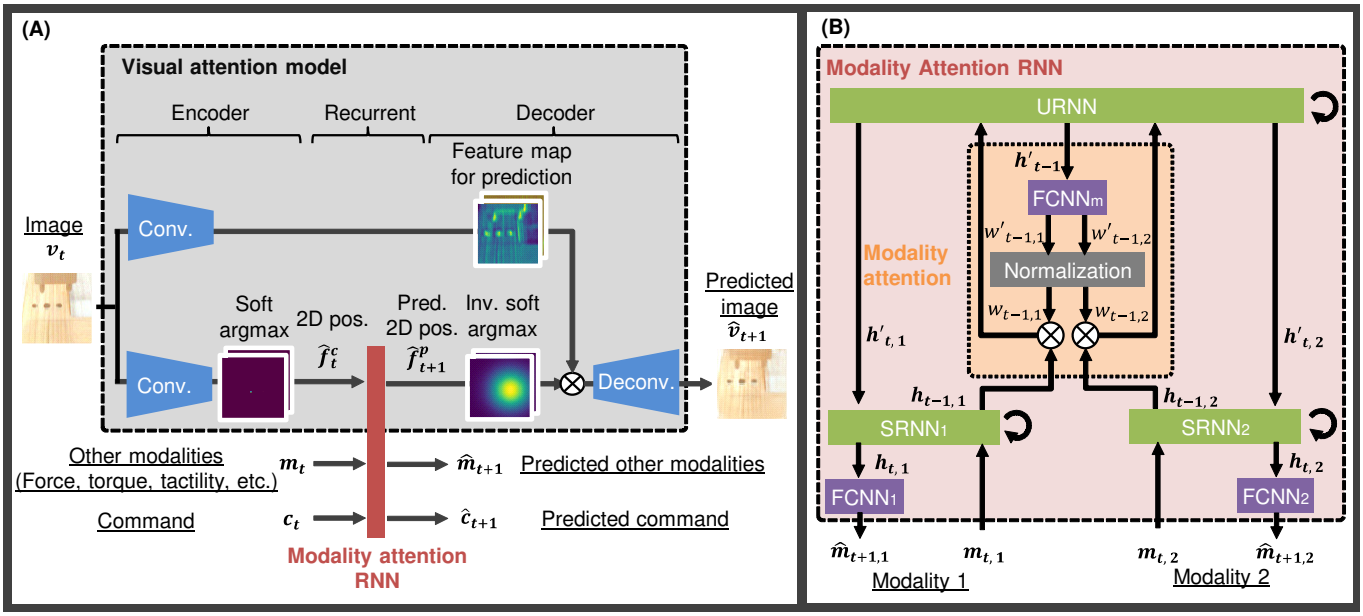

Fig. 1. Architecture of proposed modality attention model for generating robot motion. (A) Visual attention model using modality attention RNN. (B) Modality attention RNN. Separate RNNs for each modality (SRNN: Sensory RNN) and for integration (URNN: Union RNN) connected through the modality attention part by gating, and feedback from URNN to SRNN. Modality attention is obtained from the context of the URNN, and the weights are adjusted by modality predictive learning.

The model consists of three parts: an encoder, recurrent part, and decoder. The encoder extracts position coordinates in the image as image features from the camera image, the recurrent part integrates each modal and learns time-series changes, and the decoder predicts the image. The encoder utilizes the CNN and soft argmax [34] to output position coordinates $\hat{\boldsymbol{f}}_t^c$ (called attention points in this paper) of points in the image from the input image $\boldsymbol{v}_t$. On another route, feature maps are obtained using the CNN and input to the decoder. Next, in the recurrent part using fully connected NNs (FCNNs) and an RNN, the coordinates, joint angles, and other modalities at time t are used to predict those at $t+1$. The decoder predicts the image at $t+1$ on the basis of the predicted attention points and the feature maps obtained from the encoder part. Specifically, heat maps with high intensity at the predicted attention points are generated, and the image at $t+1$ is predicted from the feature map weighted by these maps. This is to support image prediction by using image features in the vicinity of the attention points. Our method utilizes RNNs that integrate multi-modality, introducing the concept of modality attention.

B. Modality attention

The proposed model is shown in Fig. 1(B). It consists of SRNNs that learn each modality in a time series, a URNN that communicates with SRNNs to learn sensory integration, and a modality attention part that gates the modalities passed to the URNN. This model extends the model of [24], which approaches differences in the physical meaning and time-varying characteristics of each modality, and introduces the concept of modality attention. There are several theories about the mechanism of attention to sensations [17]. "Efficient selection" increases sensitivity to certain sensations and decreases sensitivity to irrelevant sensory signals [17]. Inspired by [17], to achieve efficient sensory selection, sensory signals are normalized weighted between multi-modality and integrated at a higher level. Furthermore, we considered that it was important to get attention from the prediction [35], just like visual attention. Instead of creating attention from sensory signals in a bottom-up manner, our model creates attention from predictions while considering the context corresponding to the state of the task.

In Fig. 1(B), two modal information $\boldsymbol{m}_{t,1}, \boldsymbol{m}_{t,2}$ at time t is input and predict modal information $\boldsymbol{m}_{t+1,1}, \boldsymbol{m}_{t+1,2}$ at the next time $t+1$. To clarify, we assume that there are two modalities, but there may be three or more. The model consists of three FCNNs, three RNNs (SRNN, URNN), and normalization processing. Although an RNN is used, it is described as a basic network capable of learning time series, and long short-term memory (LSTM) is used in this study. The modality is predicted using Eqs. 1-5.

$$\boldsymbol{m}_{t+1,i} = \mathrm{FCNN}_i(\boldsymbol{h}_{t,i}) \quad (i=1,2), \tag{1}$$

$$\boldsymbol{h}_{t,i} = \mathrm{SRNN}_i(\boldsymbol{m}_{t,i}, \boldsymbol{h}'_{t,i}, \boldsymbol{h}_{t-1,i}), \tag{2}$$

$$\boldsymbol{w}'_{t-1} = \mathrm{FCNN}_m(\boldsymbol{h}'_t), \tag{3}$$

$$w_{t-1,i} = \frac{w'_{t-1,i}}{\sum_i w'_{t-1,i}}, \tag{4}$$

$$\boldsymbol{h}'_t = \mathrm{URNN}(w_{t-1,1}\boldsymbol{h}_{t-1,1}, w_{t-1,2}\boldsymbol{h}_{t-1,2}, \boldsymbol{h}'_{t-1}), \tag{5}$$

where $\boldsymbol{h}'_t = [\boldsymbol{h}'_{t,1}, \boldsymbol{h}'_{t,2}], \boldsymbol{w}'_t = [\boldsymbol{w}'_{t,1}, \boldsymbol{w}'_{t,2}]$. In $\mathrm{SRNN}_i (i = 1,2)$ and FCNN_i, using Eqs. 1 and 2, the modality $\boldsymbol{m}_{t+1,i}$ is predicted from the hidden state that includes the multi-modality passed from the URNN and the current modality $\boldsymbol{m}_{t,i}$. In the modality attention part highlighted in orange, each abstracted modal information $\boldsymbol{h}_{t-1,i}$ passed from SRNN_i is

weighted corresponding to the modality. Here, it is weighted by $w_{t-1,i}$ and passed to the URNN, which integrates the modalities. Since the weighting considers the context as in Eq. 3, the hidden state $\boldsymbol{h}'_{t-1}$ of the URNN one time step before is input to the FCNN, the weights $w'_{t-1,i}$ for the number of modalities are output. Note that the activation function of $FCNN_m$ is a sigmoid function to enable the output (modality weight), which is a positive value. Furthermore, by normalizing as in Eq.4, the final weight $w_{t-1,i}$ considering the signal strength of other modalities is obtained. The URNN processes each weighted modality and passes the hidden state $\boldsymbol{h}'_{t,i}$ containing multi-modality information to the modality attention part and each SRNN. Here, $w_{t,i}$ indicates how much each modality is referenced, and humans can interpret the usage of the modality by checking them. Furthermore, even if a modality with low importance is disturbed during a task, the effect of it on command prediction is reduced, so a higher robustness can be expected. Note that $w_{t,i}$ are automatically adjusted by modality predictive learning, so the teaching is not required.

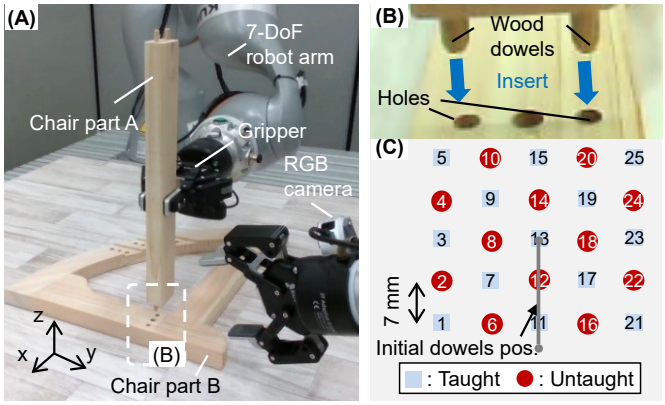

Fig. 2. Experimental conditions. (A) Overview of the experimental setup. Two robots were used, one for task execution (left) and one for camera (right). The task is to insert the wooden dowels of chair part A into the holes of chair part B placed on the table. (B) Enlarged view around the holes. Two wooden dowels are inserted into each hole. (C) Position of holes in the data set for learning and evaluation. The number of taught positions for learning is 13. The number of untaught positions for evaluation is 12.

C. Loss function

The loss function in our method is defined as follows:

$$g = \sum_{t\in T-1}(g_{t,v} + g_{t,c} + g_{t,m} + g_{t,f}), \quad (6)$$

$$g_v = \frac{1}{H\times W\times C}\|\hat{\boldsymbol{v}}_{t+1} - \boldsymbol{v}_{t+1}\|_2^2, \quad (7)$$

$$g_c = \frac{1}{M}\|\hat{\boldsymbol{c}}_{t+1} - \boldsymbol{c}_{t+1}\|_2^2, \quad (8)$$

$$g_m = \frac{1}{D}\|\hat{\boldsymbol{m}}_{t+1} - \boldsymbol{m}_{t+1}\|_2^2, \quad (9)$$

$$g_f = \frac{1}{K}\|\hat{\boldsymbol{f}}^p_{t+1} - \hat{\boldsymbol{f}}^c_{t+1}\|_2^2, \quad (10)$$

where the sequence length of the training data is T, the time is t, the image is $\boldsymbol{v} \in \mathbb{R}^{H\times W\times C}$, the joint angle $\boldsymbol{c} \in \mathbb{R}^M$ and the coordinates of the attention points $\hat{\boldsymbol{f}} \in \mathbb{R}^K$. g_v, g_c, and g_m are the prediction errors of the image, command, and other modalities , respectively. The g_f is an auxiliary loss function, which is added so that the attention points output by the recurrent part are obtained on the basis of the prediction of attention points. Note that, since the training is end-to-end, it is not necessary to teach the attention points, and g_f is calculated between the outputs of the NN.

IV. Experiments

A. Experimental setup

Fig. 2(A) shows the experimental setup. Robot arms are KUKA LBR iiwa 7 R1400, and the gripper is a Robotiq 2F-85 Adaptive Gripper. Additionally, a camera (Intel RealSense L515) was placed at the tip of the robot arm on the right side of Fig. 2(A) for the robot's vision. For the experimental task, we selected the task of inserting furniture parts (IKONIH's kids chair) as one that requires multi-modality. The purpose of the task is to insert the two wooden dowels of part A into the two holes of part B, as shown in Fig. 2(B). This task includes the flow of general tasks (pick and place, door opening, wiping, etc.) of reaching, contacting, and interacting with the work object to accomplish the task. Therefore, this task was selected from the viewpoint of its generality as well as considering the ease of evaluating interpretability. The tasks were performed in the order shown in Fig. 3. Note that "pre-approach" is the contact motion of parts A and B to visually grasp the relative height of the parts, and is not the motion of approaching the wooden dowel into the hole. Furthermore, it is a task that should refer to vision when "approaching" and force when "inserting". The motion data of the robot when the initial posture of the robot and part A were fixed and the position of part B was changed, was used as learning data. Fig. 2(C) shows the position of part B in the xy plane when acquiring learning data (blue □) and the untaught position (red ◯). The small ∘ in the figure indicates the initial position of the wooden dowel of part A. The figure shows only the position of the right hole shown in Fig. 2(B), with 13 taught positions for learning and 12 untaught positions for verification experiments set as intermediate positions between the taught points. The number shown in the hole position indicates the index of the position. Data is obtained twice for each taught position, and a total of 26 data are used for learning. These data were collected by manipulating the robot arm using the end-effector pose control of the arm using a game controller, and acquiring the joint angle data of the robot arm at 10 Hz, the image data of the camera, and the force data for 15 seconds. Impedance control was used to control the robot to reduce the risk of damage to objects and robots.

The purpose of the experiment was to verify the applicability, interpretability, and robustness of the modality attention model to robot tasks. In this study, we focused on overfitting for specific modalities when dealing with multi-modality. We introduced a disturbance by testing an extreme case where a modality is completely useless. Disturbances were added to a modality, and the timings were estimated to be of low importance during task execution. For this purpose, we learned three models: (i) a model w/o modality attention and force,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

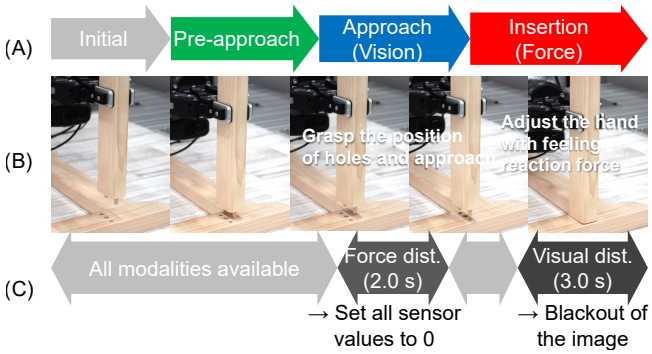

Fig. 3. Order of task execution. (A) Task state. Modalities expected to be important are described in parentheses. (B) Snapshots. (C) Disturbance added during the verification experiment.

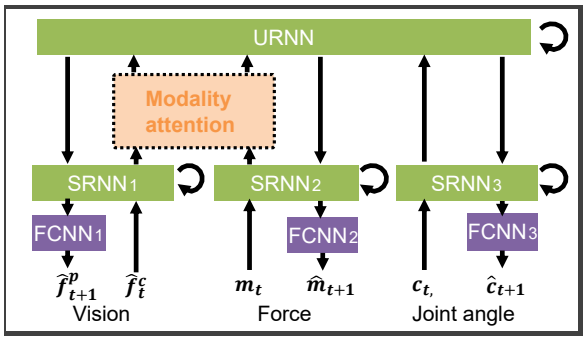

Fig. 4. Modality attention RNN architecture in the experiments. There are three modalities: camera image as vision, end-effector force estimated from the torque sensors as force, and joint angle as the command.

(ii) a model w/o modality attention and w/ vision and force, and (iii) a model w/ modality attention, vision and force. (i) is to show that this task requires multi-modality, and the task success rate is evaluated under non-disturbance. (ii) and (iii) were used to examine the effectiveness of modality attention and evaluated task success rates in three cases: non-disturbance, with force disturbances during "Approach," and with visual disturbances during "Insertion," as shown in Fig. 3(C). In this study, we used the modality attention RNN model shown in Fig. 4. The modalities are vision, force, and joint angles, and the joint angles are not gated. This is because joint angles have always been considered an important modality, as they are equal to the robot's motion commands. Gating the joint angles is considered effective when forces and torques are treated directly as command values. In each case, 25 positions of part B including untaught positions shown in Fig. 2(C) were tested three times for each position for a total of 75 times, and the success rate was evaluated. The robot motion time was set to 15 seconds, which is the same as the sequence length of the learning data. These disturbances are situations of out-of-distribution, unlike the situation when learning. It was set assuming unexpected camera occlusion and temporary sensor error in the real environment. Table I shows the network parameters of the models. Additionally, the batch size was 13, the input data was scaled to [0,1.0], and 12,000 epochs were learned. For training, the Adam optimizer [36] with parameters $\alpha = 0.001, \beta_1 = 0.9$, and $\beta_2 = 0.999$

TABLE I
EXPERIMENT HYPERPARAMETERS.

	Layer type	Parameter	Output shape
Encoder	CNN 1_1	k=3, s=1, c=16	62×62×16
	CNN 1_2	k=3, s=1, c=32	60×60×16
	CNN 1_3	k=3, s=1, c=8	58×58×8
	Soft argmax	-	8×2
	CNN 2_1	k=3, s=1, c=16	62×62×16
	CNN 2_2	k=3, s=1, c=32	60×60×16
	CNN 2_3	k=3, s=1, c=8	58×58×8
Recurrent	SRNN	n=50	50
	FCNN*	n=2×8, 7, 6, 2	2×8, 7, 6, 2
	URNN	n=20	20
Decoder	Inv soft argmax	-	58×58×8
	Transposed CNN 1	k=3, s=1, c=32	60×60×32
	Transposed CNN 2	k=3, s=1, c=16	62×62×16
	Transposed CNN 3	k=3, s=1, c=3	64×64×3

c, k, s, and n denote the number of channels, kernel, strides and output dimensions (nodes), respectively.
* Three FCNNs in the non-modality attention model (one for each modality). For the modality attention, another FCNN (FCNN_m) is required.

TABLE II
TASK SUCCESS RATE FOR EACH MODEL AND CASE.

Model \ Task	Non-dist. (95% CI)	Visual dist. (95% CI)	Force dist. (95% CI)
(i) Vision only	28.0 ± 10.0	-	-
(ii) Non-modality att.	88.0 ± 7.4	45.3 ± 11.0	41.3 ± 10.9
(iii) Modality att.: ours	**90.7 ± 6.7**	**88.0 ± 7.4**	**89.3 ± 7.1**

(default Tensorflow settings) was used. Furthermore, the AI Bridging Cloud Infrastructure system of the National Institute of Advanced Industrial Science and Technology was used. The system has 16 GB per node and we used one node. The time for the learning was about 4 hours for both models.

B. Results

1) Without disturbance: Table II shows the task success rate for each model and case. For vision only, the success rate is significantly lower than the case in which force is also used. This suggests that both vision and force are important for performing this task. Furthermore, the task success rate under non-disturbance was 88.3% and 90.6% with the non-modality and modality attention models, respectively, and the success rate with modality attention was slightly higher. However, considering the variance, it cannot be said that there is a significant difference. It was found that gating modality information by modality attention did not adversely affect the success rate without disturbances. Fig. 5 shows (A) input images and attention points, (B) predicted images, (C) bird's-eye view images, and (D) the transition of modality attention for the modality attention model. Time elapses in the image on the right. In (A), there are two types of attention points: × indicates current attention, and ○ indicates predicted attention. Attention points are appropriately obtained at positions important for the task, such as part A with wooden dowels and holes in part B. In Fig. 5 (D), the red dashed and blue solid lines represent the attention to vision and to force, respectively. During "Pre-approach," force changed due to the contact with

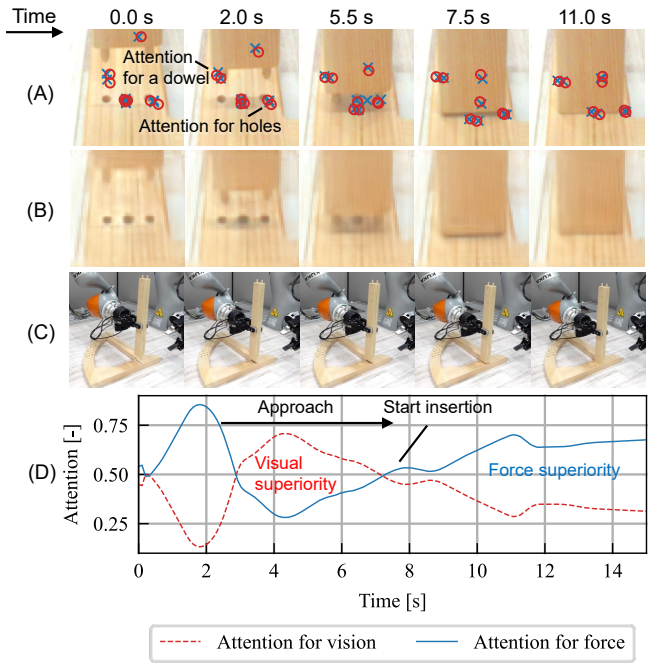

Fig. 5. Snapshots of the chair part insertion task. (A) Input images with attention points(×: current attention points, ◯: predicted attention points). (B) Predicted images. (C) Bird's-eye view images (not for control). (D) Transition of modality attention.

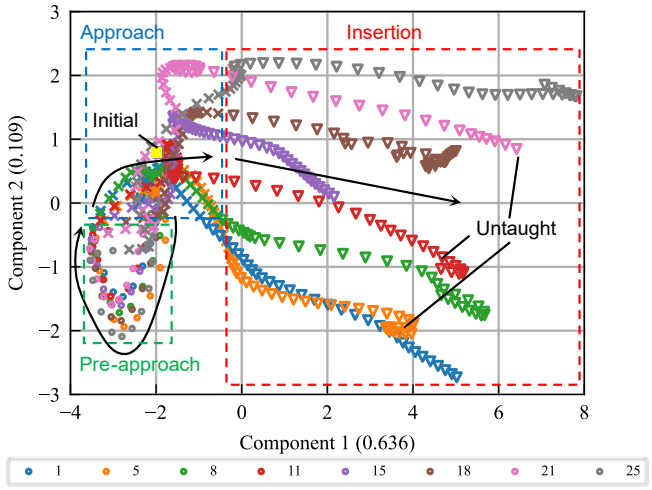

Fig. 6. Principle component analysis (PCA) results for hidden states of the URNN in the proposed model based on the position of part B. The first and second components are shown. A total of eight points are plotted for each position of the part. It is structured by task state and part position.

the parts, so the change was predicted and attention was directed to force. During "Approach," it is necessary to grasp the position of the hole on the basis of vision, so attention was directed to vision. During "Insertion," attention was directed to force to recognize the insertion state of the hole. It also matches human intuition. In this way, people can check the modality attention, and it can be said that it contributes to the improvement of interpretability.

The expected role of the URNN in processing gated modalities is to retain critical information while filtering out unnecessary information for task execution. The former is

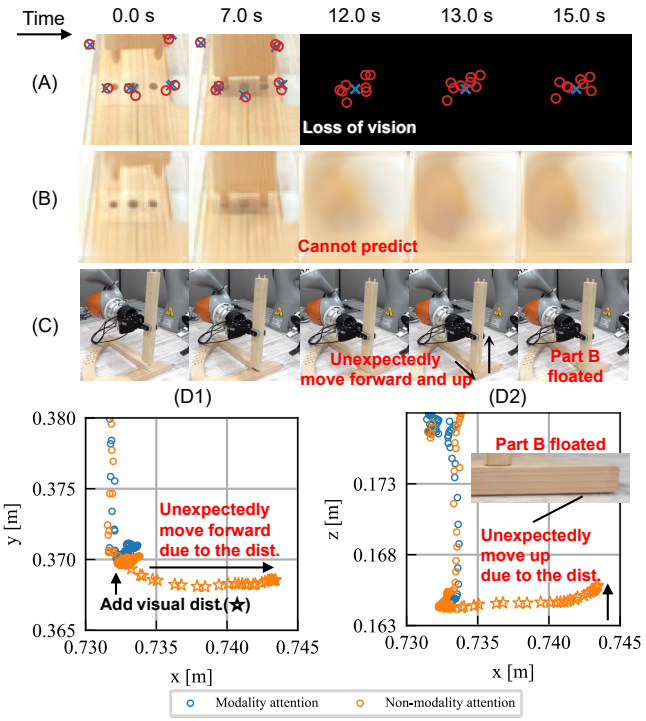

Fig. 7. Snapshots of the task with the non-modality attention under visual disturbance. (A) Input images with attention points (×: current attention points, ◯: predicted attention points). (B) Predicted images. (C) Bird's-eye view images(not for control). (D1) Transition of the end-effector position of (xy coordinate). (D2) Transition of the end-effector position (xz coordinate).

confirmed by how modality attention was obtained. The latter is considered through PCA of hidden states of the URNN, as shown in Fig. 6. The color is changed for each position of part B and plotted. The initial time is indicated by yellow □, ○ for "Pre-approach," × for "Approach," and ▽ for "Insertion" corresponding to the task state. A different area is plotted for each task state and for each position of part B. In particular, for the position of part B, the case of the untaught position is plotted between the taught positions, indicating proper self-organization. Although the modality attention model gates information, it was found to be able to achieve both robustness and task execution performance because the architecture is capable of extracting important information for task, such as the task state and object position.

2) With disturbance: Here, we describe the case under visual and force disturbances, respectively. The modality attention model maintained a high success rate even with disturbances. This is because, in the same way as shown in Fig. 5, attention is focused on vision during "Approach" and force during "Insertion", regardless of the state of any disturbance. However, in the non-modality attention model, the success rate was significantly lower under disturbances than without disturbances. Fig. 7 shows snapshots of the task with the non-modality attention model under visual disturbance. (D1) and (D2) show the comparison of the transition of the end effector position in the motion between the modality attention model (success case) and the non-modality attention model (failure case). (D1) and (D2) indicate the xy and xz coordinates, respectively. ⋆ indicates the timing when the

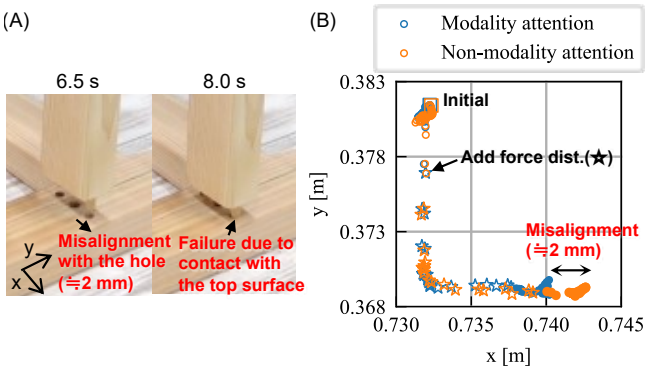

Fig. 8. Motions of the non-modality attention model under force disturbance. (A) Snapshots of the mistake. (B) Transition of the end-effector position (xy coordinate).

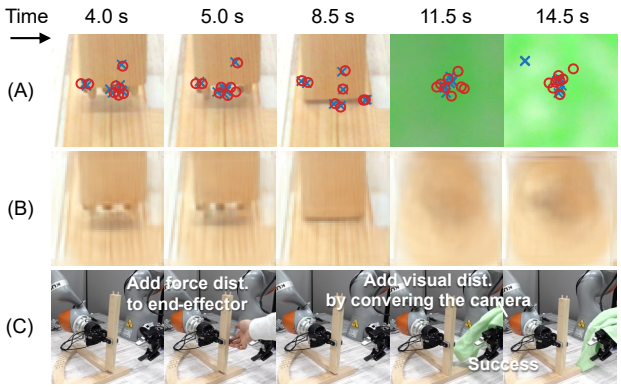

Fig. 9. Snapshots of the chair part insertion task with actual possible force and visual disturbances. (A) Input images with attention points(×: current attention points, ◯: predicted attention points). (B) Predicted images. (C) Bird's-eye view images (not for control).

disturbance was added. At 12.0 s, the vision was lost, so attention was focused on the center, and the prediction of the image was not possible. Furthermore, the robot unexpectedly moved forward (+x) and upward (+z) when the disturbance was applied. This action lifted part B and the condition is undesirable because lifting the part while it is partially inserted puts a large load on the part. Figures 8 (A) and (B) show the motions of the non-modality attention model under force disturbance and the transition of the end-effector position compared between the modality attention model (success case) and the non-modality attention model (failure case). In the non-modality attention model, the end position of "Approach" was shifted forward (+x) from the robot compared with the modality attention model, and the robot contacted the upper surface of the part when inserting it, failing the task. Note that the non-modality attention model succeeded in the task at this part position under non-disturbance, indicating that the disturbances adversely affected the joint angle prediction of the model. From the aforementioned comparison, it was found that the robustness against disturbances can be improved by switching the modality of attention on the basis of prediction.

C. Demonstration

We confirmed the motion of the modality attention model when possible disturbances were added to the robot. As a force disturbance, a human hand applied force to the end-effector, and as a visual disturbance, the camera was covered. Furthermore, to verify applicability to other tasks, we applied the model to the outlet plug insertion task and marker wiping task. Several results have been omitted due to space limitations, so please refer to the attached video.

1) With real disturbance: Figure 9 shows snapshots of the chair part insertion task. We confirmed that the modality attention model worked effectively and succeeded in the task under both types of disturbances. Similar experiments were also done for the non-modality attention model, but failed.

2) Other application: Snapshots for the outlet plug insertion and marker wiping tasks are shown in Fig. 10. In the outlet plug insertion task, the position of the strip was fixed, and the height deviation of the plug was learned. In the marker wiping task, the deviation of the marker position and board height

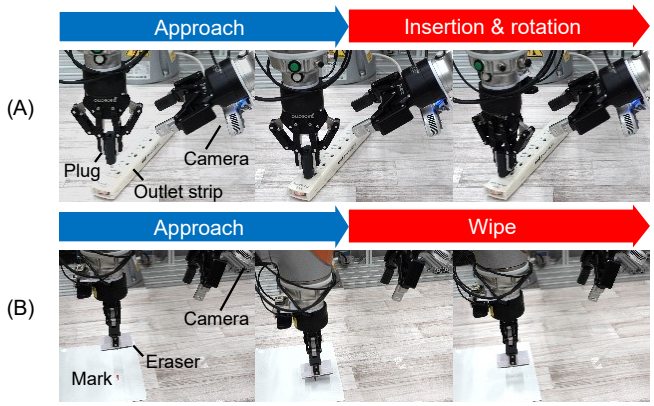

Fig. 10. Example of other applications. (A) Outlet plug insertion task. (B) Marker wiping task.

was learned. In both cases, the transition of modality attention was confirmed, and the robustness against disturbance was confirmed.

V. Conclusion

We proposed a modality attention model on the basis of multi-modality prediction. This was achieved by gating each modality to the integrated RNN in the model consisting of two types of RNN, one that processes each modality in a distributed manner and one that integrates it. When it was verified in a chair part insertion task using vision and force, it was confirmed that vision was dominant when approaching and force was dominant when inserting. In contrast to the non-modality attention model, whose task success rate drops significantly under disturbance, the modality attention model enhances robustness against disturbances, resulting in a consistently high success rate. In the future, we aim to make robots work autonomously under unpredictable disturbances, such as when working with humans.

Acknowledgement

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology (AIST) was used.

REFERENCES

[1] E. Macaluso and J. Driver, "Multisensory spatial interactions: a window onto functional integration in the human brain," *Trends in neurosciences*, vol. 28, no. 5, pp. 264–271, 2005.

[2] C. Finn, X. Y. Tan, Y. Duan, T. Darrell, S. Levine, and P. Abbeel, "Deep spatial autoencoders for visuomotor learning," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 512–519.

[3] Y. Tsurumine, Y. Cui, E. Uchibe, and T. Matsubara, "Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation," *Robotics and Autonomous Systems*, vol. 112, pp. 72–83, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889018303245

[4] H. van Hoof, N. Chen, M. Karl, P. van der Smagt, and J. Peters, "Stable reinforcement learning with autoencoders for tactile and visual data," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 3928–3934.

[5] K. Noda, H. Arie, Y. Suga, and T. Ogata, "Multimodal integration learning of robot behavior using deep neural networks," *Robotics and Autonomous Systems*, vol. 62, no. 6, pp. 721–736, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889014000396

[6] T. Adachi, K. Fujimoto, S. Sakaino, and T. Tsuji, "Imitation learning for object manipulation based on position/force information using bilateral control," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3648–3653.

[7] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Transformer-based deep imitation learning for dual-arm robot manipulation," *CoRR*, vol. abs/2108.00385, 2021. [Online]. Available: https://arxiv.org/abs/2108.00385

[8] M. Okada and T. Taniguchi, "Dreaming: Model-based reinforcement learning by latent imagination without reconstruction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4209–4215.

[9] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "Daydreamer: World models for physical robot learning," *arXiv preprint arXiv:2206.14176*, 2022.

[10] H. R. Walke, J. H. Yang, A. Yu, A. Kumar, J. Orbik, A. Singh, and S. Levine, "Don't start from scratch: Leveraging prior data to automate robotic reinforcement learning," in *Conference on Robot Learning*. PMLR, 2023, pp. 1652–1662.

[11] M. Vecerik, O. Sushkov, D. Barker, T. Rothörl, T. Hester, and J. Scholz, "A practical approach to insertion with variable socket position using deep reinforcement learning," in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 754–760.

[12] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Contact-rich manipulation of a flexible object based on deep predictive learning using vision and tactility," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 5375–5381.

[13] H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Efficient multitask learning with an embodied predictive model for door opening and entry with whole-body control," *Science Robotics*, vol. 7, no. 65, p. eaax8177, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aax8177

[14] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Spatial attention point network for deep-learning-based robust autonomous robot motion generation," *arXiv preprint arXiv:2103.01598*, 2021.

[15] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," *arXiv preprint arXiv:2302.02408*, 2023.

[16] H. Hiruma, H. Ito, H. Mori, and T. Ogata, "Deep active visual attention for real-time robot motion generation: Emergence of tool-body assimilation and adaptive tool-use," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8550–8557, 2022.

[17] F. Pestilli, M. Carrasco, D. J. Heeger, and J. L. Gardner, "Attentional enhancement via selection and pooling of early sensory responses in human visual cortex," *Neuron*, vol. 72, no. 5, pp. 832–846, 2011.

[18] A. Mandlekar, D. Xu, R. Martín-Martín, S. Savarese, and L. Fei-Fei, "Learning to generalize across long-horizon tasks from human demonstrations," *arXiv preprint arXiv:2003.06085*, 2020.

[19] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397–403, 2016.

[20] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 734–743.

[21] H. Ito, T. Kurata, and T. Ogata, "Sensory-motor learning for simultaneous control of motion and force: Generating rubbing motion against uneven object," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 408–415.

[22] K. Kawaharazuka, A. Miki, M. Bando, K. Okada, and M. Inaba, "Dynamic cloth manipulation considering variable stiffness and material change using deep predictive model with parametric bias," *Frontiers in Neurorobotics*, vol. 16, 2022.

[23] N. Saito, T. Shimizu, T. Ogata, and S. Sugano, "Utilization of image/force/tactile sensor data for object-shape-oriented manipulation: Wiping objects with turning back motions and occlusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 968–975, 2022.

[24] H. Ichiwara, H. Ito, K. Yamamoto, H. Mori, and T. Ogata, "Multimodal time series learning of robots based on distributed and integrated modalities: Verification with a simulator and actual robots," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9551–9557.

[25] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8943–8950.

[26] Y. Liu, D. Romeres, D. K. Jha, and D. Nikovski, "Understanding multimodal perception using behavioral cloning for peg-in-a-hole insertion tasks," *arXiv preprint arXiv:2007.11646*, 2020.

[27] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Transformer-based deep imitation learning for dual-arm robot manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8965–8972.

[28] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[29] S. Cui, R. Wang, J. Wei, J. Hu, and S. Wang, "Self-attention based visual-tactile fusion learning for predicting grasp outcomes," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5827–5834, 2020.

[30] T. Anzai and K. Takahashi, "Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9361–9368.

[31] M. A. Lee, M. Tan, Y. Zhu, and J. Bohg, "Detect, reject, correct: Cross-modal compensation of corrupted sensors," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 909–916.

[32] K. Lee, Z. Wang, B. Vlahov, H. Brar, and E. A. Theodorou, "Ensemble bayesian decision making with redundant deep perceptual control policies," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 831–837.

[33] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: bottom-up versus top-down," *Current biology*, vol. 14, no. 19, pp. R850–R852, 2004.

[34] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5137–5146.

[35] Y.-F. Hsu, J. Hamalainen, and F. Waszak, "Both attention and prediction are necessary for adaptive neuronal tuning in sensory processing," *Frontiers in Human Neuroscience*, vol. 8, 2014. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00152

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.