2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Conditional NewtonianVAE to Generate Pre-Grasping Actions in Physical Latent Spaces

Masaki Ito[1], Gustavo Alfonso Garcia Ricardez[1,*], Ryo Okumura[2], and Tadahiro Taniguchi[1]

***Abstract*— To make robotic grasping scalable, vision-based control with high data efficiency and accuracy is needed. World models are capable of creating representations of physical environments from sensory information. In particular, NewtonianVAE is a world model that can control targets in physical environments by using proportional control in its latent space from input images. However, NewtonianVAE entangles information of each object in separate state subspaces making control unfeasible when trained with multiple objects. In this paper, we introduce Conditional NewtonianVAE, a novel framework designed to generate pre-grasping actions by disentangling object-type information from the state space in physical latent spaces. Our method incorporates a conditioning variable to achieve disentanglement, facilitating the use of the learned state space for control tasks. Through simulation and real-robot experiments, we demonstrate the effectiveness of Conditional NewtonianVAE in accurately positioning the end-effector into a pre-grasping pose, thereby enhancing the success rate of robotic grasping. Conditional NewtonianVAE achieves a grasping success rate of 83% for known objects and 78% for unseen objects in the real-robot experiments.**

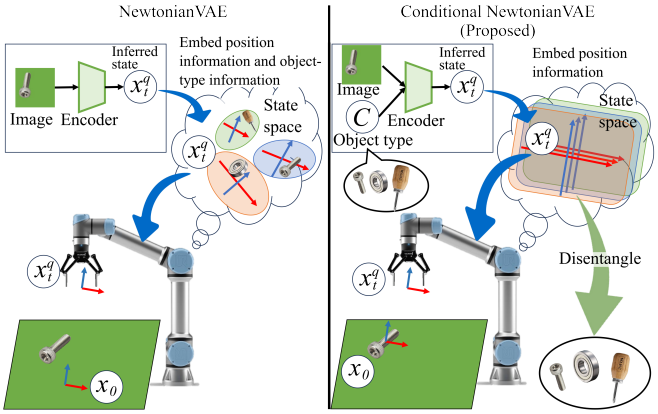

Fig. 1. Right: the proposed method can disentangle the object-type information from the state space by adding a conditioning variable, enabling the state space to be used for positioning tasks for multiple objects in the dataset. Left: Without disentanglement, each trained object has its own subspace, rendering positioning tasks unfeasible.

I. Introduction

With the world-widespread need for robotic automation to counteract the labor shortage caused by aging societies, creating methodologies to boost the capabilities of robots to manipulate various objects has become crucial. Typically, robotic grasping is realized following the *perception-then-action* paradigm which requires processing visual information and then planning to grasp objects [1], [2]. Nevertheless, these state-of-the-art methods are costly to train and are decoupled from control.

World models [3], [4] can create a representation (i.e., state space) of the external physical environment (i.e., action space) from sensory information such as images [5], [6], [7]. In particular, NewtonianVAE is a type of world model that acquires a latent space equivalent to mapping from images to physical representations [8]. It is designed to induce proportional controllability in the latent space, enabling the use of proportional control and goal identification from pixels via physical latent spaces. By enabling simple control from images, NewtonianVAE removes complex learning of control policies, which results in high data efficiency.

However, when learning representations for multiple objects, NewtonianVAE entangles the information of each object in separate state subspaces, which renders control unfeasible. Therefore, a mechanism is needed to prevent the embedding of object-type information into the state space.

In this paper, we propose to add a conditioning variable into NewtonianVAE to disentangle the object-type information, which enables using the learned state space for control tasks. In our work, we aim to enable robotic grasping by setting the pre-grasping action as a positioning task, i.e., controlling the end-effector so that the target object is at a pre-defined position within the camera view. We assume that a successful positioning task leads to a pre-grasping pose from which grasping is feasible. In our foreseen settings, the camera is attached to the end effector next to the gripper. We verify in simulation and real-robot experiments the positioning accuracy and the grasping success rate of the proposed method, respectively.

Our main contributions are:

1) We propose a general framework to disentangle object-type information from the state space, which enables control tasks in the action space when training NewtonianVAE with multiple objects.
2) We demonstrate the effectiveness of the proposed method to position the end-effector into a pre-grasping pose by verifying the grasping success rate in real-robot experiments.

This work was supported by the Japan Science and Technology Agency (JST), Moonshot Research & Development Program, Grant Number JPMJMS2011.

[1]Masaki Ito, Gustavo Alfonso Garcia Ricardez, and Tadahiro Taniguchi are with Ritsumeikan University; 1-1-1 Noji-Higashi, Kusatsu, Shiga 525-8577, Japan. {garcia-g, taniguchi}@em.ci.ritsumei.ac.jp

[2]Ryo Okumura is with Panasonic Corporation; 1006 Kadoma, Kadoma, Osaka 571-0050, Japan. okumura.ryo001@jp.panasonic.com

*Corresponding author.

The remainder of this paper is organized as follows. We provide an overview of related works in Section II. We detail the integration of the proposed system in Section III. We describe our experimental results in Section IV and provide a discussion in Section V. Finally, we conclude and add directions for future work in Section VI.

II. Related Works

A world model is a type of model that can acquire a state space that represents the external environment [3], [4]. Various methods for generating agent behaviors based on the world model have been proposed [9], [10], [11], [12], [13], [14]. In general, motion generation can be achieved by combining a world model with policy learning. Therefore, it is necessary to perform multiple interactions with the environment and collect a considerable amount of training data.

NewtonianVAE is a kind of world model that can control an object using a state space obtained by inputting an image to an encoder as a visual motion [8]. Therefore, it is possible to control the target using only the world model, eliminating the need to learn complex policies. The core of NewtonianVAE is a dynamical model based on Newton's equations of motion. A high correlation between the actual environment and the state space allows NewtonianVAE to control the target to a desired position. Due to its simple learning of control, NewtonianVAE has fast convergence, which makes it highly data efficient.

Tactile-sensitive NewtonianVAE is a method based on NewtonianVAE for modeling the state of grasping a USB by tactile sensors [15]. When the robot grasps the USB, there is always an error in the grasping position of the USB (i.e., the gripper not always grasps the USB at the same grasping points), so when inserting the USB, the insertion position must be adjusted to account for the error in the grasping position. For this purpose, the RGB image of the USB being grasped is acquired using the tactile sensor GelSight [16], and the insertion position adjusted with the error in the grasping position is embedded in the state space. By controlling the robot with the adjusted insertion position as a target, the USB can be moved to the exact insertion position. Their work shows that NewtonianVAE achieves a high precision when realizing an action and a high success rate. Similarly, Kato et al. proposed in hand-view-sensitive Newtonian variational autoencoder (ihVS NVAE), a method that uses images of grasped objects to estimate their pose, while the objects are still being grasped, and they applied it to a box-packing task [17].

Our proposed method is inspired by the work of Sohn et al. who extended the variational autoencoder (VAE) [18], [19] by adding a conditioning variable and proposed the Conditional VAE [20] method. In this paper, we aim to separate the object-type information from the state space by adding a conditioning variable.

III. Proposed Method

In this section, we describe Conditional NewtonianVAE, which is a model that introduces a conditioning variable into NewtonianVAE. By adding a conditioning variable, our proposed method can disentangle the object information from the representation and achieve a generalized state space. The obtained state space has a high correlation with the external environment and, therefore, can be used for control tasks such as centering a target object in the camera's view.

More specifically, Conditional NewtonianVAE is a method for constructing a state space with controllable positioning when using multiple types of objects by introducing conditioning variables to NewtonianVAE. The conditioning variable can be seen as a bias toward the state space and the reconstructed image. In other words, while learning is encouraged to embed features that best represent the image when only an image is input, the introduction of conditioning variables makes it possible to encourage learning to embed features other than the content of the image since the content of the image is explicitly given. Though conditioning variables can have different representations, we choose to use one-hot vectors for simplicity. This is because it is easy to assign the type of control object to the elements of the vector, although the length of the vector becomes longer as the number of control object types increases.

Conditional NewtonianVAE consists of the following three distribution functions:

$$q(\mathbf{x}_t|\mathbf{I}_t, \mathbf{C}), \tag{1}$$

$$p(\mathbf{I}_{t+1}|\mathbf{x}_{t+1}, \mathbf{C}), \tag{2}$$

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t; \mathbf{v}_t), \tag{3}$$

where $\mathbf{x}_t$ is the state space, $\mathbf{I}_t$ is the image information, $\mathbf{C}$ is the conditioning variable, $\mathbf{v}_t$ is the velocity, and $\mathbf{u}_t$ is the robot action. By maximizing the evidence lower bound (ELBO) of the Conditional NewtonianVAE, it is possible to learn a controllable state space for the control target. The ELBO of the Conditional NewtonianVAE is as follows:

$$\begin{aligned}\mathcal{L} = \mathbb{E}_{q(\mathbf{x}_t|\mathbf{I}_t,\mathbf{C})q(\mathbf{x}_{t+1}|\mathbf{I}_{t+1},\mathbf{C})} \\ \Big[\mathbb{E}_{p(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{u}_t;\mathbf{v}_t)}[p(\mathbf{I}_{t+1}|\mathbf{x}_{t+1},\mathbf{C})] \\ - \mathrm{KL}(q(\mathbf{x}_{t+1}|\mathbf{I}_{t+1},\mathbf{C}) \mid\mid p(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{u}_t;\mathbf{v}_t))\Big],\end{aligned} \tag{4}$$

where, $q(\mathbf{x}_t|\mathbf{I}_t, \mathbf{C})$, $p(\mathbf{I}_t|\mathbf{x}_t, \mathbf{C})$ and $p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t; \mathbf{v}_t)$ assume a Gaussian distribution.

When maximizing Eq. (4), it is possible to learn a state space that is highly correlated with the coordinate system of the external environment. A high absolute correlation means that the direction of change of the position inferred from the input image in the state space is the same as the direction of change of the object's position in the external environment. In other words, the encoder should be able to output consistent position information as the object's position in the external environment changes.

Finally, Conditional NewtonianVAE can be used for robot control using the obtained state space. Eq. (5) is used to

Fig. 2. Objects from the YCB set used in the simulation experiments. The position of the object within the camera view is used as a) the starting position for the random walks during the data collection process (i.e., all random walks start with the object in the same position within the camera view), and b) the target position of the positioning task (i.e., the methods should control the end effector until the object is in such position within the camera view).

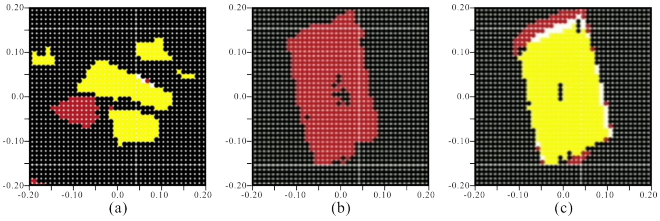

Fig. 3. Visualization of the state space (a) NewtonianVAE, (b) Conditional NewtonianVAE with condition $(1, 0)$, and (c) Conditional NewtonianVAE with condition $(0, 1)$.

calculate the control quantity:

$$\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{I}_t), \tag{5}$$

$$\mathbf{u}_t = K_p(\mathbf{x}_t^{goal} - \mathbf{x}_t), \tag{6}$$

where $\mathbf{x}_t$ is the position information in the state space, $q(\mathbf{x}_t|\mathbf{I}_t)$ is the approximate distribution, K_p is the gain of the proportional controller, and $\mathbf{u}_t$ is the control quantity.

IV. Experimental Results

We perform three experiments to evaluate the proposed method to i) verify the disentanglement of the object-type information and the state space, ii) evaluate the accuracy of the control task, and iii) measure the grasping success rate. The experiments (i) and (ii) are realized in simulation while (iii) is realized with a real robot.

A. Disentaglement Verification

We prepare a simulation environment using MuJoCo [21] with the objects (1) and (2) of Fig. 2, and a green background. The objects are placed inside the virtual camera view. To gather data, we start with each object at the same position, and then randomly move the virtual camera (i.e., realize a random walk) while keeping the object within the camera view. We collect the image information $\mathbf{I}$, control variable $\mathbf{u}$, and object position; the latter is used to calculate the correlation coefficient.

After training, we visualize the state space by traversing it within a given range and color-coding it according to the objects in the reconstructed image, as shown in Fig. 3. We represent the object (1) as red, object (2) as yellow, and *no object* as black. Fig. 3(a) shows the visualization of the state space of NewtonianVAE, while Figs. 3(b) and (c) that of the proposed Conditional NewtonianVAE, where one-hot vectors of $(1, 0)$ and $(0, 1)$ are input as conditioning information, respectively.

As shown in Fig. 3(a), we confirmed that the object type information is embedded in the state space, as each object creates its own subspace. On the other hand, Figs. 3(b) and (c) show that the introduction of the condition information can separate the object type information from the state space. In other words, this indicates that explicitly inputting condition information can be used to ensure that only location information is embedded in the state space.

TABLE I
Correlation Between the Position of the State Space (Latent) and Action Space (Real)

Model	Correlation	
	X	Y
NewtonianVAE	0.829	0.500
Conditional NewtonianVAE (Proposed)	**0.897**	**0.722**

A comparison of the correlation coefficients between the position in the state space and the position in the action space of the control target in the simulation environment is shown in Table I. The object-type information is embedded in the state space obtained by training NewtonianVAE which reduces the absolute value of the correlation coefficient. On the other hand, Conditional NewtonianVAE can separate the object-type information from the state space by introducing condition information. Therefore, the only information embedded in the state space is location information. As a result, the absolute value of the correlation coefficient of Conditional NewtonianVAE is higher than that of NewtonianVAE.

B. Accuracy Evaluation

We use the prepared simulation environment with all objects shown in Fig. 2, and collect data in the same way explained previously. We evaluate the accuracy by measuring the RMSE between the target position of the object (i.e., the centered position used when creating the dataset) and the position achieved by the proposed and compared method. We call this *positioning task*. Fig. 4 shows an example of the positioning task using Conditional NewtonianVAE.

Moreover, we quantify the number of successes in the positioning task. We deem the positioning task successful when the error is less than 70 mm, which is motivated by our ultimate goal of grasping objects with a 2-finger gripper whose maximum width is 140 mm; the distance for the TCP to one extended finger would be 70 mm, hence the 70-mm threshold.

We compare the proposed method to the original NewtonianVAE, which allows us to investigate the impact on the accuracy when using state spaces with and without the separation of the object-type information by explicitly inputting the condition information.

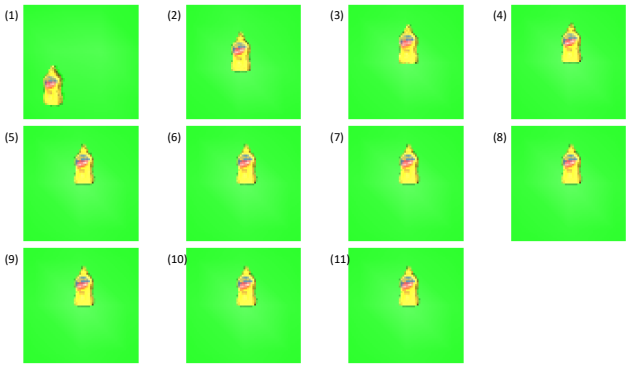

Fig. 4. Positioning task with Conditional NewtonianVAE. The object is placed at a random location within the camera view and Conditional NewtonianVAE controls the end effector above the object until the object is at the pre-defined location used during training.

TABLE II
POSITIONING ACCURACY OF CONDITIONAL NEWTONIANVAE IN SIMULATION

Object	NewtonianVAE		Conditional NewtonianVAE (Proposed)	
	Successes		Successes	
	#	Error ($\times 10^{-3}$)	#	Error ($\times 10^{-3}$)
Mustard bottle	3/50	53.85 ± 12.76	**50/50**	$\mathbf{1.30 \pm 0.00}$
Tomato soup can	0/50	-	**50/50**	$\mathbf{2.19 \pm 0.00}$
Cracker box	0/50	-	**50/50**	$\mathbf{0.25 \pm 0.00}$
Master chef can	2/50	35.61 ± 7.35	**50/50**	$\mathbf{0.95 \pm 0.00}$
Gelatin box	0/50	-	**50/50**	$\mathbf{0.85 \pm 0.00}$
Pudding box	0/50	-	**50/50**	$\mathbf{0.75 \pm 0.01}$
Sugar box	0/50	-	**50/50**	$\mathbf{0.85 \pm 0.00}$
Tuna fish can	0/50	-	**50/50**	$\mathbf{1.97 \pm 0.00}$
Potted meat can	0/50	-	**45/50**	$\mathbf{2.18 \pm 0.00}$
Power drill	0/50	-	**50/50**	$\mathbf{0.87 \pm 0.00}$
	5/500		**495/500**	

The evaluation of the accuracy for each object in the positioning task with 10 different object types is shown in Table II. We found that for all object types, Conditional NewtonianVAE increases the accuracy in the positioning task compared to NewtonianVAE. With Conditional NewtonianVAE, we confirmed that the error in positioning each object is small. Moreover, only five failures were observed, which corresponded to the Potted meat can.

The results indicate that Conditional NewtonianVAE is more suitable for the positioning task, as its error is smaller compared to NewtonianVAE. This can be attributed to the fact that, by providing the condition variable explicitly, only the location information is embedded in the state space, and there is no object-specific location information as in the vanilla NewtonianVAE.

C. Grasping Success Rate

We set up an object grasping task as a positioning task in a real-world environment and verify how the state spaces obtained with Conditional NewtonianVAE and NewtonianVAE affect the number of successful grasp attempts. Furthermore, we evaluate how the proposed method generalizes to unseen objects by providing conditioning information of similar objects.

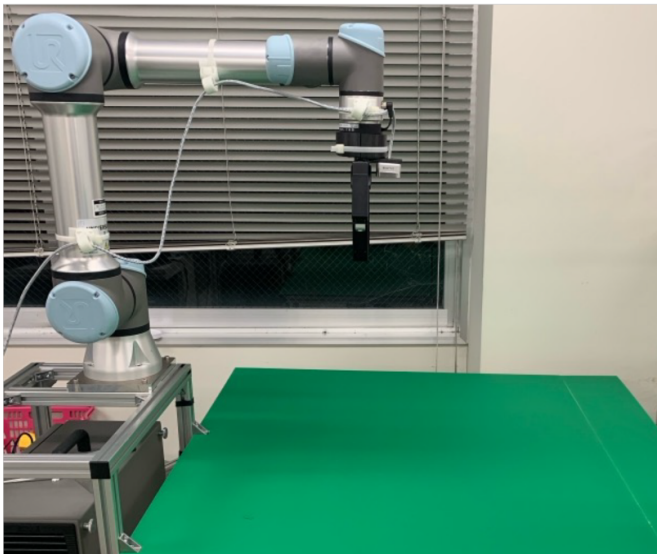

Fig. 5. Testbed for real-world experiments. A manipulator UR5e with a 2-finger Robotiq gripper and a RealSense D405 camera.

Fig. 6. Objects used for the training dataset. From right to left and top to bottom: Mustard bottle, Chocolate box, String, Tomato soup can, Baseball, Sugar box, Pringles, Cracker box, Screwdriver, and Lock.

We prepare the experimental setup shown in Fig. 5, which consists of a manipulator with a camera and a 2-finger gripper, and a green table where objects are placed. We select 10 objects from the YCB set for the training dataset and select five objects not included in the dataset nor part of the YCB set as unseen objects, which are shown in Figs. 6 and 7, respectively. To gather data, we use a similar approach to the simulation case. The data is collected in 30 episodes per object, and 15 images are collected per episode. Therefore, 450 images are collected per object. The size of the collected images is 64×64 pixels. For each method, a single model is trained using the data of all objects (training dataset).

Each trial starts with the object in a random location within the camera view, and the proposed or compared method controls the end effector so that the object is at a location within the camera view corresponding to the trained location, at the same height as in the dataset. Once the corresponding

Fig. 7. Unseen objects used in the experiments to test the generalization capabilities of the proposed method. From right to left: Tennis ball, Chip star, Pringles B, Soup can, and Cookie box. These objects were not included in the training dataset. As conditioning variable, we set the one corresponding to a similar object in the training dataset.

TABLE III
GRASPING SUCCESS RATE OF KNOWN AND UNSEEN OBJECTS

Object	NewtonianVAE	Conditional NewtonianVAE	
	Successes	Condition	Successes
Known Objects			
Mustard bottle	0/20	Mustard bottle	**18/20**
Chocolate box	0/20	Chocolate box	**16/20**
String	0/20	String	**19/20**
Tomato soup can	0/20	Tomato soup can	**17/20**
Baseball	0/20	Baseball	**6/20**
Sugar box	0/20	Sugar box	**20/20**
Pringles	0/20	Pringles	**18/20**
Cracker box	0/20	Cracker box	**19/20**
Screwdriver	0/20	Screwdriver	**16/20**
Lock	0/20	Lock	**17/20**
	0/200		**166/200**
Unseen Objects			
Tennis ball	-	Baseball	**1/20**
Chip star	-	Pringles	**20/20**
Pringles B	-	Pringles	**19/20**
Soup can	-	Tomato soup can	**20/20**
Cookie box	-	Tomato soup can	**18/20**
			78/100
	0/200		**244/300**

method has finished controlling the end effector for that purpose, the end effector is lowered directly over the z-axis for a predefined distance, relative to the height of the dataset. Finally, the gripper closes the fingers to grasp the object, and the end effector moves directly up over the z-axis. If the object is successfully grasped, the trial is deemed successful. In the case of Conditional NewtonianVAE, a one-hot vector corresponding to the target object is provided as condition information; if an unseen object is targeted, the one-hot vector corresponding to a similar object chosen from the training dataset is provided. Fig. 8 shows an example of Conditional NewtonianVAE used for grasping, including the camera view and the reconstruction.

The results with **known objects** are shown in Table III. NewtonianVAE failed to grasp the object, whereas Conditional NewtonianVAE was able to grasp the objects 166 times out of 200 trials (83%). This indicates that Conditional NewtonianVAE can acquire a state space that can be used for a positioning task in the real world.

The results with **unseen objects** are shown in Table III. The Pringles B and Chip star are labeled as Pringles, the Tennis ball is labeled as Baseball, and the Cookie box and Soup can are labeled as Tomato soup can. The number of successful grasps of Pringles B and Chip star was similar to that of the Pringles used in the training dataset. The number of successful grasps of the Cookie box and the Soup can was similar to that of the Tomato soup can used in the training dataset. These results suggest that the encoder of Conditional NewtonianVAE may be robust to differences in size and texture of the same object type. Though it was difficult to grasp the Tennis ball, this is the same result as the grasping success rate of the Baseball used in the training dataset.

V. DISCUSSION

The number of successful grasping of the Baseball by Conditional NewtonianVAE was low. This may be due to the fact that the target positions of the Baseball may have pulled to the target positions of objects other than the Baseball. Although the conditioning variable is effective in removing information on the object type from the state space, it is difficult to strongly capture the object-specific target positions for positioning. Therefore, we believe that the target position was pulled by the similar-looking String, which affected the number of successful grasp attempts of the Baseball.

The target position of the positioning task is determined by that set in the dataset used for training. As the dataset considers a position above the object from where grasping is judged feasible, Conditional NewtonianVAE consequently learns also the pre-grasping positions. As the generalization of the proposed method is high, the pre-grasping positions can be used for grasping unseen objects as long as the selected condition (e.g., one-hot vector $(0, 1)$) corresponds to an object with similar features, as shown in Table III. However, the inaccuracies in the object used as condition will be inherited, as in the case of the Tennis ball.

Though the proposed method can separate the object-type information from the state space, such separation is sensitive to overlapped features such as similar colors in the dataset, as shown in Fig. 3(c), where red letters on the Mustard bottle are encoded as part of the red-ish Tomato soup can.

VI. CONCLUSION

We proposed a general framework in which Conditional NewtonianVAE can embed positional information in the state space independent of object type. Simulation and real-world experiments have shown that Conditional NewtonianVAE is capable of locating the origin or grasping position in the external environment for multiple types of objects. In addition, we set up an object grasping task in a real-world environment and found that the number of successful grasp attempts was higher than that of NewtonianVAE. Furthermore, when we performed grasping on unseen objects with different textures and sizes, we found that the number of successful grasping attempts was similar to that of trained objects.

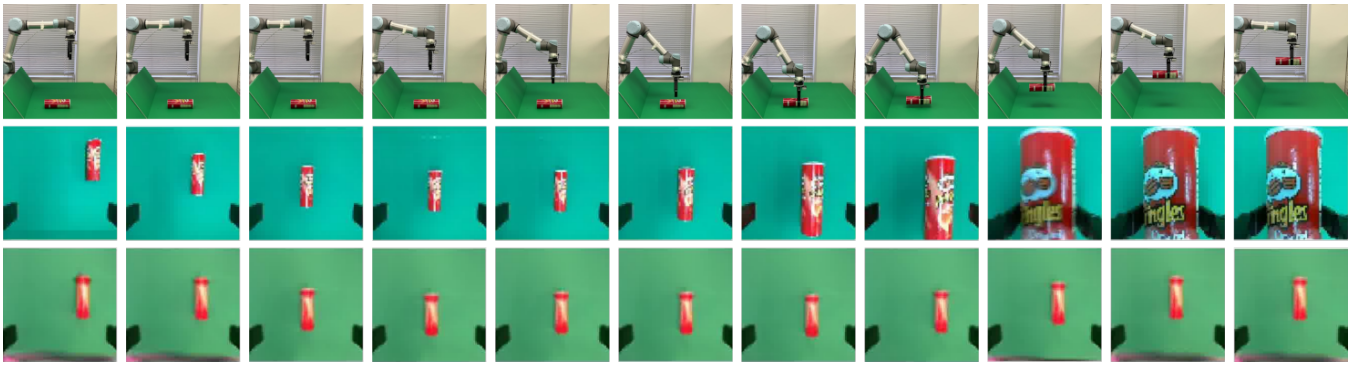

Fig. 8. Object grasping in the real world with Conditional NewtonianVAE. Top: robot action to grasp an object. Middle: input image from the end effector's camera. Bottom: reconstruction by the proposed method. Note that the training dataset contains images at a fixed height, which is why the size of the reconstructed images (bottom) corresponds to such height, even when the end effector approaches the object and grasps it.

In future work, we plan to investigate the impact on the number of successful positioning attempts and the error from the target position by utilizing the features obtained from image encoders of models such as CLIP [22], as object-type information. This will enable us to examine whether it is possible to embed location information into the state space independent of the number of object types. Therefore, the application of the model to environments with a larger number of object types can be considered.

In this work, we focused on the positioning task in the xy-plane, while the z-axis and all rotations are not considered. In particular, the equivariant, invariant, and covariant rotations are not explicitly encoded by CNNs and, therefore, neglected when training the proposed method. A method that includes these rotations and the z-axis changes should also be investigated.

REFERENCES

[1] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[2] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "AnyGrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[3] K. Friston, R. J. Moran, Y. Nagai, T. Taniguchi, H. Gomi, and J. Tenenbaum, "World model learning and inference," *Neural Networks*, vol. 144, pp. 573–590, 2021.

[4] T. Taniguchi, S. Murata, M. Suzuki, D. Ognibene, P. Lanillos, E. Ugur, L. Jamone, T. Nakamura, A. Ciria, B. Lara, *et al.*, "World models and predictive coding for cognitive and developmental robotics: Frontiers and challenges," *Advanced Robotics*, vol. 37, no. 13, pp. 780–806, 2023.

[5] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[6] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[7] A. Richard, S. Aravecchia, M. Geist, and C. Pradalier, "Learning behaviors through physics-driven latent imagination," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 1190–1199.

[8] M. Jaques, M. Burke, and T. M. Hospedales, "NewtonianVAE: Proportional control and goal identification from pixels via physical latent spaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4454–4463.

[9] R. S. Sutton, "Integrated modeling and control based on reinforcement learning and dynamic programming," in *Advances in Neural Information Processing Systems*, R. Lippmann, J. Moody, and D. Touretzky, Eds., vol. 3. Morgan-Kaufmann, 1990.

[10] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," *Advances in neural information processing systems*, vol. 31, 2018.

[11] M. Okada, N. Kosaka, and T. Taniguchi, "PlaNet of the Bayesians: Reconsidering and improving deep planning network by incorporating Bayesian inference," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5611–5618.

[12] 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.

[13] M. Okada and T. Taniguchi, "Dreamingv2: Reinforcement learning with discrete world models without reconstruction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 985–991.

[14] A. Kinose, M. Okada, R. Okumura, and T. Taniguchi, "Multi-view dreaming: Multi-view world model with contrastive learning," *Advanced Robotics*, vol. 37, no. 19, pp. 1212–1220, 2023.

[15] R. Okumura, N. Nishio, and T. Taniguchi, "Tactile-sensitive NewtonianVAE for high-accuracy industrial connector insertion," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4625–4631.

[16] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 818–824.

[17] Y. Kato, R. Okumura, and T. Taniguchi, "World-model-based control for industrial box-packing of multiple objects using NewtonianVAE," *arXiv preprint arXiv:2308.02136*, 2023.

[18] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[19] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[20] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[21] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 5026–5033.

[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.