2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

InterRep: A Visual Interaction Representation for Robotic Grasping

Yu Cui[1], Qi Ye[1†], Qingtao Liu[1], Anjun Chen[1], Gaofeng Li[1], and Jiming Chen[1]

Abstract— **Recently, pre-trained vision models have gained significant attention in motor control, showcasing impressive performance across diverse robotic learning tasks. While previous works predominantly concentrate on the significance of the pre-training phase, the equally important task of extracting more effective representations based on existing pre-trained visual models remains unexplored. To better leverage the representation capabilities of pre-trained models for robotic grasping, we propose InterRep, a novel interaction representation method that possesses not only the strengths of pre-trained models, known for their robustness in noisy environments and their proficiency in recognizing essential features, but also the capacity of capturing dynamic interaction details and local geometric features during the grasping process. Based on the novel representation, we introduce a deep reinforcement learning method to learn generalizable grasping policies. The experimental results demonstrate that our proposed representation outperforms the baselines in terms of both training speed and generalization. For the generalized grasping tasks with dexterous robotic hands, our method boasts a success rate nearly 20% higher than methods using the global features of the entire image from pre-trained models. In addition, our proposed representation method demonstrates promising performance when applied to a different robotic hand and task. It also exhibits excellent performance on real robots with a success rate of 70%.**

I. Introduction

The pre-training paradigm has yielded encouraging outcomes in computer vision [1], [2] and natural language processing [3], [4]. Recently, an expanding realm of research has been dedicated to establishing pre-trained models as the fundamental building blocks for enhancing policy learning in vision-based motor control [5], [6]. Given that these models are usually trained on extensive datasets of real-world images, their features encompass a broad understanding of the semantics and characteristics of our world. The flourishing of pre-training methodologies promises to revolutionize the field by imbuing agents with a deep understanding of the visual world, enabling them to excel in various downstream tasks.

However, most recent research in leveraging pre-training for robot learning solely focuses on how to train a better pre-trained model itself, such as incorporating masked autoencoder (MAE) [7] for pre-training [8], [9] or integrating language [10], [11]. When deploying the pre-trained model in downstream robot learning tasks, previous works primarily use the latent visual representation of the entire image, without delving further into the exploration of more effective representations.

Towards better harnessing the representation capabilities of pre-trained models, we incorporate the dynamic interaction information into the representation. Prior research [12], [13] has demonstrated that extracting the dynamic interaction cues in the whole manipulation process from images can assist robots in better understanding the underlying structure of scene tasks and gaining a more precise knowledge of object conditions and environmental alterations, leading to enhancing generalization abilities. VIOLA [12] decomposes a visual scene into a set of factorized representations of objects extracted from raw images, while GraphIRL [13] abstracts agent-object interactions via a graph representation. However, these two interaction representation methods mentioned above either ignore object shape information or rely on cropping the entire object region, posing challenges in achieving satisfactory performance and generalization to new objects, especially for grasping tasks involving complex-shaped objects or dexterous grasping with rich contacts.

To leverage the advantages of visual pre-trained models and enhance their generalization performance in grasping, we introduce a novel visual representation named **InterRep**. Concretely, at each step during the grasping process, we first use a pre-trained model as our visual encoder to extract the features of the entire image. Then we integrate dynamic interaction information by selecting local features of the regions closest to the robotic hand on the object. This process focuses on the areas most likely to come into contact with the robotic hand, capturing distance information between the hand and the object, as well as the object's local shape details. The method we propose helps improve generalization to new objects, because, despite their different overall shapes, these objects often share similar geometric features in local regions or are composed of similar basic geometric shapes. For example, various teapots and cups may have similarities in their handle shapes.

To validate the effectiveness and generalization capability of our proposed representation method, we integrate it into a reinforcement learning framework and conduct experiments on dexterous grasping tasks with Adroit Hand [14]. The results show our method can successfully grasp various objects with different shapes and exhibit a strong generalization ability to novel objects, achieving an almost 20% increase in success rate compared to the existing method using the global feature of the entire image. Furthermore, it is validated

[1]College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China.

[†]Qi Ye (Corresponding author, qi.ye@zju.edu.cn) is with the College of Control Science and Engineering, the State Key Laboratory of Industrial Control Technology, Zhejiang University, and the Key Key Lab of CS&AUS of Zhejiang Province.

This work was supported in part by the National Natural Science Foundation of China (Grant Number: 62088101, 62103372, 62233013).

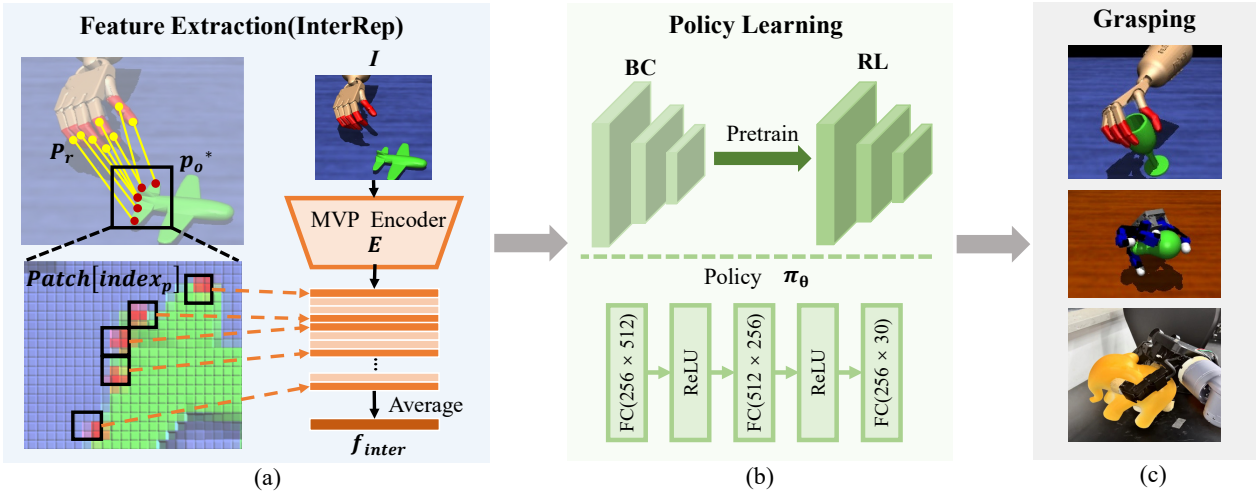


Fig. 1: **Overview of Our Method.** (a) shows our proposed representation InterRep, (b) shows the framework of policy learning, and (c) shows our grasping tasks.

to be effective on real robots.

Our main contributions are listed as follows:

- We propose a novel visual representation for robot grasping, combining the strengths of pre-trained vision models with the dynamic interaction characteristics during the grasping process.
- Our proposed method outperforms baselines using the latent visual representation of the entire image in both simulation and the real world. Experiments also demonstrate its effectiveness in different experimental setups.

II. Related Work

Visual Representation Learning for Robotics. Recent developments in the field of robot learning focus on acquiring visual state representations for control. Some prior methodologies relied on extracting representations from domain-specific data directly sourced from the target environment and its associated tasks [12], [13], [15]. In some cases, this extraction process may require object detection [16], [17] to identify and extract objects. With the advancements of pre-training in computer vision [1], [2], another area of research focuses on harnessing representations obtained from these pre-trained models. In particular, Parisi et al. [6] evaluate several pre-trained vision representations trained with supervised or self-supervised learning on various robot learning tasks, demonstrating promising results. Nair et al. [5] leverage egocentric video data to pre-train visual representations, enhancing performance in robotic manipulation tasks. Additionally, Radosavovic et al. [8] demonstrate that pre-training with MAE [7] on large-scale videos has also proven effective. Unlike the aforementioned work, our approach combines the advantages of both task-specific information and pre-trained models.

Policy Learning. Reinforcement learning (RL) [18] and imitation learning (IL) [19] represent two predominant paradigms in policy learning, achieving remarkable outcomes across a broad spectrum of control tasks. For High-degree-of-freedom manipulation tasks, many works combine RL and IL to enhance sample efficiency. Among these, DAPG [20] augments policy gradients with expert demonstrations in reinforcement learning. Additionally, ILAD [21] proposes a novel imitation learning approach, which utilizes expert demonstrations to expedite the reinforcement learning training phase. In this work, we focus on learning visual-motor policies based on DAPG [20] for both dexterous and simple manipulation policy learning.

III. Method

In this section, we first introduce our proposed novel interaction representation that possesses both the strengths of pre-trained models and the capacity to capture dynamic interaction details and local geometric features during the grasping process. Following that, we introduce the RL policy learning framework incorporating our proposed features. This framework is based on DAPG (Demonstration Augmented Policy Gradient) [20], a two-stage policy learning approach comprising behavior cloning for initialization and reinforcement learning for fine-tuning. Our pipeline is illustrated in Fig. 1.

A. Problem Formulation

When addressing the challenge of robotic grasping, we conceptualize it as a discrete-time Markov Decision Process (MDP). The standard MDP is formally characterized by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces respectively. $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ represents the transition dynamics, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the reward function, and $\gamma \in (0, 1]$ signifies the discount factor. The goal is to maximize the reward with a policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$.

B. Masked Visual Pre-Training (MVP)

We employ the vision encoder in MVP [8], which is a representative and one of the state-of-the-art pre-trained vision models for robotic manipulation. MVP is pre-trained on images from diverse in-the-wild videos via MAE [7], making it well-suited for real-world robotic tasks. Considering that

we focus more on the manipulation tasks which consist of close-range interaction between objects and robotic hands, we choose the ViT-S HoI model referred to as E trained on the Hand-object Interaction (HoI) data used in [9].

C. Interaction Representation

This section describes how to build our proposed interaction representation. We first obtain the regions of the object and the robotic hand by segmenting the masks, and then compute the interaction features.

1) Object Region Extraction: In a scenario where a robot interacts with an object, the area of greatest concern for the robot is the area where the interacting object is located. Therefore, we first extract the regions of the objects. For simple scenarios, we segment the object directly via colors. For scenes with more complex backgrounds, we leverage the advantages of large visual models. Specifically, we first render an initial frame of scenario I_0 and utilize SAM model $\mathcal{F}_{sam}$ [22] to segment the object via point prompts. Then the mask predicted by SAM is fed into the Xmem model $\mathcal{F}_{xmem}$ [23] which can track the mask step by step during the RL training process. Finally, the object mask of each frame M_o is extracted as:

$$M_o = \mathcal{F}_{xmem}\left(\mathcal{F}_{sam}\left(I_0\right)\right). \tag{1}$$

2) Interaction Features Computing: Since it is necessary to focus on the dynamic interaction process between the robotic hand and the object during manipulation, apart from segmenting the object, we also apply the same method as Sec. III-C.1 to obtain the robotic hand mask denoted as M_r in the RGB image to facilitate subsequent feature extraction.

In the feature computing stage, We use E as the vision encoder and keep it frozen during RL policy training. Taking each frame of the manipulation task scene I as input, the model outputs the embedding feature computed as:

$$f = E\left(I\right), f \in \mathbb{R}^{197\times 384}. \tag{2}$$

For the global feature f_g of the current frame, only the $0th$ dimension of the output needs to be taken. For the local interaction features we proposed, we sample m pixels in M_r and n pixels in M_o, and then respectively record the coordinates of the selected pixels in the original image, denoted as P_r and P_o. In order to get the dynamic local features of the hand-object interaction process, the distance between P_r and P_o is computed, and the pixels of the object closest to the hand denoted as ${p_o}^*$ are selected:

$${p_o}^* = argmin\left(\left\|P_o, P_r\right\|_2\right). \tag{3}$$

Then the k corresponding patch indexes $\left\{(index_p)_{i=0}^k\right\}$ to which the pixels ${p_o}^*$ belong in the whole image are computed. We acquire local interaction features $f_{inter} \in \mathbb{R}^{384}$ by extracting the features corresponding to these k patches from f and calculating their average:

$$f_{inter} = average(f[(index_p)_{i=0}^k]). \tag{4}$$

Particularly, when no robotic hand or object is present in one frame, the global feature f_g of this frame is utilized.

D. Policy Learning Framework

1) Demonstration Retargeting: Following DexRepNet [24], we obtain demonstrations from GRAB [25]. For each human trajectory in GRAB, We only extract the right-hand data and capture the sequence from a distance of 10 cm from the object to a height of 4 cm when lifted off the table. In order to retarget human demonstrations to robotic grasping trajectories, we follow the method of [26] to formulate the retargeting objective as a nonlinear optimization problem with the cost function:

$$\min_{q^{\mathbf{R}},T} \sum_{i=0}^{N} \left\|\mathbf{v_i^R}\left(T, q^{\mathbf{R}}\right) - k_i \mathbf{v_i^H}\left(q^{\mathbf{H}}\right)\right\|_2, \tag{5}$$

where $\mathbf{vi^R}$ and $\mathbf{vi^H}$ are distance vectors between fingers and between the hand and the object, computed through forward kinematics using the joint angles $q^{\mathbf{R}}$ and $q^{\mathbf{H}}$ respectively. k_i is the scale ratio between the robotic hand and the MANO Hand [27]. Additionally, the global transformations T of the robotic arm are incorporated into the optimization process, resulting in improved performance. After optimization, We convert the optimized joint angles into actions within the MuJoCo [28] environment. Subsequently, we execute the fine-tuned action sequences employed correlated sampling [29] in a simulation setting to gather demonstrations $\mathcal{D}$ consisting of state-action pairs (s, a), for behavior cloning in the subsequent phase.

2) Behavior Cloning (BC): We follow prior works [20], [21] to leverage BC to reduce the sample complexity of training an RL policy from scratch. With $\mathcal{D}$, we pretrain the policy π_θ with the objective:

$$\mathcal{L}_{bc} = \sum_{(s,a)\in\mathcal{D}} \left\|\pi_\theta(s) - a\right\|_2, \tag{6}$$

where θ is the parameter of the policy network to be optimized.

3) RL Training with InterRep: Although using BC to initialize can accelerate training speed, it doesn't fully utilize demonstration data and constrained exploratory behavior. DAPG [20] uses RL to fine-tune the BC policy and adds an additional term to the gradient:

$$\begin{aligned} g_{\text{aug}} = &\sum_{(s,a)\in\rho_\pi} \nabla_\theta \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s,a) + \\ &\sum_{(s,a)\in\mathcal{D}} \nabla_\theta \ln \pi_\theta(a \mid s) \lambda_0 \lambda_1^k \max_{(s,a)\in\rho_\pi} A^{\pi_\theta}(s,a), \end{aligned} \tag{7}$$

where ρ_π represents the trajectories sampled by policy π_θ, and $\mathcal{D}$ corresponds to the collected demonstration data. A^{π_θ} is the advantage function, and k is the iteration counter. λ_0, λ_1 are hyper-parameters set to 0.1 and 0.95 respectively.

Observation Space. In addition to InterRep mentioned above, we incorporate proprioception, denoted as f_{pro}, into the observation input. f_{pro} exclusively includes the joint angles of the Adroit hand. Before being fed into the policy network, f_{pro} and InterRep f_{inter} are respectively processed through a single-layer fully-connected encoder, which has

the output size of 128 and accepts an input size of 30 for f_{pro} or 384 for f_{inter}. Consequently, The final observation is represented as:

$$O = [FC(f_{pro}), FC(f_{inter}))]. \tag{8}$$

The weights of these fully connected layers are frozen after behavior cloning.

Action Space. We employ the Adroit platform [14] as our manipulator, comprising a 6-DoF arm paired with a 24-DoF hand. Consequently, the action space corresponds to the continuous motor commands of 30 actuators.

Reward Function. To promote grasping effectiveness, our reward function contains three components:

$$R = \alpha R_{dis} + \beta R_{lift} + \gamma R_{succ}, \tag{9}$$

where $\alpha = 0.1$ and $\beta = \gamma = 1$. R_{dis} is a reward related to the closest distance between the fingertips of the robot hand and the surface of the object, encouraging the hand to approach the object. R_{lift} is the lifting reward when the object is lifted off the table. R_{succ} is a bonus constant reward when the object's position reaches the target position. The calculation details of the rewards can be referred to in [24].

IV. Experiments

In this section, we conduct a series of experiments to answer the following questions:

1) Can our method grasp diverse objects and generalize well to unseen objects?
2) Can our method be applicable to different experimental setups?
3) Is our method practical for real-world deployment?

A. Experimental Setup

Simulation Environments. We construct our grasping simulation environment in MuJoCo [28] simulator. In the initial step, we position each object at the origin of the world coordinate system and set the target grasp position $0.15m$ above the object. To introduce variability, we apply a disturbance along the x-axis within the range of $[-0.05m, 0.05m]$ and along the y-axis within the range of $[-0.05m, 0]$. Additionally, we randomly rotate the object around the z-axis within the range of $[-\pi, \pi]$. The object's physical properties are defined as follows: it has a mass of 0.5kg and a friction coefficient of 1. For rendering, a camera is fixed in the third-person view and renders a new image every five steps. Furthermore, we implement an initialization step where we set the hand to the average pose observed in the initial steps of all retargeted demonstrations, which also ensures a natural pre-grasp pose.

Object Dataset. We validate our approach with two object datasets: GRAB [25] and 3DNet [30]. Due to the limitations of the simulator in collision detection, we generate collision meshes for each object mesh by employing convex decomposition with CoACD [31], and scale the objects to fit the size of the robotic hands. For policy training, we train a unified policy encompassing all 40 objects from the GRAB dataset. Our evaluation begins by assessing the grasp performance on the left 10 objects in GRAB [25]. Subsequently, we proceed to test our model's capability on a set of 28 novel object meshes obtained from 3DNet [30].

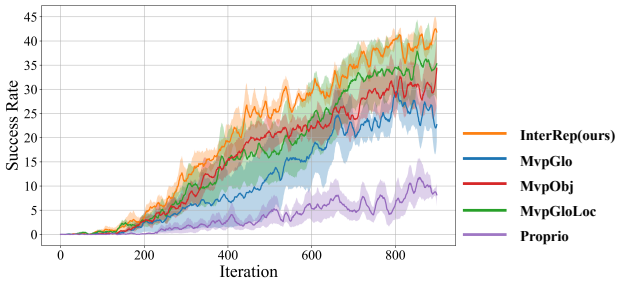


Fig. 2: Success rates (%) of our method and baselines during RL training.

Metrics. We employ the success rate of grasping as the primary evaluation metric to gauge the effectiveness of our approach. Specifically, within a single episode, we define a successful grasp as one where an object is securely held within a proximity of $3cm$ from the target position for a minimum of 50 steps.

Implementation Details. We train the grasping policy, which is a 3-layer MLP with ReLU activations between each layer, on the device with Intel Xeon Gold 6326 and NVIDIA 3090. In the phase of BC, all parameters are optimized using the Adam optimizer for 200 epochs with a minibatch size of 64 and a learning rate of $1e-5$. In the phase of RL training, the grasping policy is trained for about 900 iterations with 3 random seeds. At each iteration, every object is sampled for 2 trajectories in the environment, resulting in a cumulative total of 80 exploration-generated trajectories, and the length of each trajectory is 200 steps. When evaluating, each object is sampled for 10 trajectories, and the final success rate is computed as the average across all objects.

B. Baseline

1) **Proprio:** This baseline only uses proprioceptive states (joint positions) of the robotic hand.
2) **MvpGlo:** This baseline uses both **Proprio** and the global feature f_g extracted from MVP encoder E as described in Section III-B.
3) **MvpObj:** This baseline combines **Proprio** and the features of all pixels of the object segmented from the image, without any dynamic interaction information between the robotic hand and the object.
4) **MvpGloLoc:** This baseline combines **Proprio**, f_g, and the interaction image feature f_{inter} to investigate whether the integration of our approach with the global feature of the entire image is more effective, i.e. **MvpGlo+InterRep**.

C. Results

1) Efficiency and Generalization Ability of InterRep: Fig. 2 illustrates the training progress of dexterous grasping with the Adroit Hand. Compared to the baselines, our approach demonstrates the fastest convergence, achieving the highest success rate.

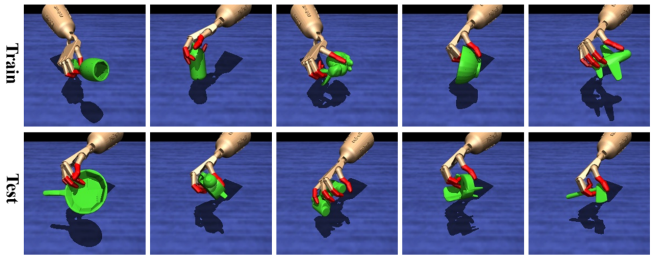
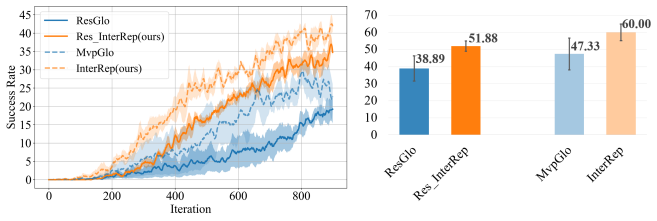


Fig. 3: **Grasping Visualization.** The first row is objects for training and the second row is objects for testing.



(a) Training on 40 objects of GRAB (b) Test on 10 unseen objects of GRAB

Fig. 4: Comparison of success rates (%) between our method (orange) and the method using global features (blue) on different pre-trained models.

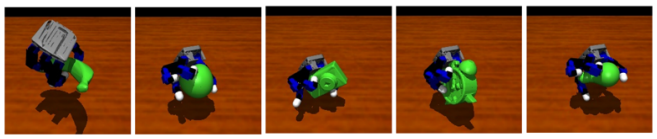


Fig. 5: Effectiveness of grasping with Allegro in simulation.

To validate the generalization ability of our method on novel objects, we evaluate the policy after 900 training iterations on unseen objects from GRAB [25] and 3DNet [30]. The result is the average of three random seeds, shown in Tab. I. Our method can still achieve a success rate of over 60% and outperforms other baselines. The results suggest that our method excels in capturing the shared features across diverse objects, enabling effective generalization to novel objects.

Impact of Dynamic Interaction Information. Comparing our method **InterRep** with **MvpGlo**, we find extracting representations related to the dynamic interaction between the hand and the object is effective for enhancing the success rate and generalization of grasping. From the perspective of human manipulation of objects, describing a visual scene in terms of objects and their interactions enables humans to quickly learn and make precise predictions [15]. Consequently, introducing interaction information can enable robots to better mimic human actions, enhancing the naturalness and efficiency of task execution. Moreover, dynamic interaction information helps robots understand the state of objects and changes in the environment more accurately and better understand the underlying structure of scene tasks, which can improve generalization abilities.

TABLE I: Success rates (%) of grasping with Adroit tested on 10 unseen objects of GRAB and 28 objects of 3DNet.

Methods	GRAB	3DNet	Average
MvpGlo [8]	47.30	42.00	44.68
Proprio	25.00	23.56	24.28
MvpObj	49.33	54.03	51.28
MvpGloLoc	**61.30**	63.00	62.15
InterRep(ours)	60.00	**66.30**	**63.15**

Impact of Local Geometry Feature. The results of comparing our method **InterRep** with **MvpObj** suggest that compared with directly extracting the features of all pixels of the object, extracting only the features of local pixels of the object can improve training efficiency and generalization ability. The reason why local features exhibit better generalization is twofold: firstly, local features are typically more robust to variations. When certain parts of an object or scene change, local features remain relatively more stable. Secondly, even though their overall shapes may vary significantly for different categories of objects, their local regions usually share similar curves or basic patterns. In addition, the local features are computed based on the dynamic interaction between the robotic hand and the object, encompassing crucial semantic information about the interaction during manipulation. Such information aids the model in a more comprehensive and rapid understanding of the image content.

Impact of Object Region Extraction. Comparing **MvpObj** with **MvpGlo**, we find extracting the features of all pixels of the object still yields better results than taking the entire image feature directly. This is likely because, in the field of robot learning, the features of the manipulated object often contain crucial task-related information, assisting robots in better comprehending and learning the essence of the task more effectively. This emphasis helps reduce interference from irrelevant information, ultimately enhancing the efficiency and performance of the learning process.

Impact of the Combination of Global and Local Interaction Features. Comparing **MvpGloLoc** with our method **InterRep** in Fig. 2 and Tab. I, we find that the integration of global features does not yield an improvement in performance; on the contrary, it leads to a deterioration in performance during the training process. This could be due to information redundancy between the two types of features. Moreover, for grasping tasks that prioritize details, the background information from the entire image may not be very useful and could even introduce interference.

2) Application to Different Pre-Trained Models: This section aims to validate the effectiveness of **InterRep** across different pre-trained models. In Fig. 4, the two solid lines represent the MVP pre-trained model used in our main experiments, while the two dashed lines represent ResNet50 [32] pre-trained on ImageNet. It can be observed from the figure that our approach remains effective when using a pre-trained model with a different architecture.

3) Application to Different Robotic Hands and Tasks: This section validates the generalization of **InterRep** in

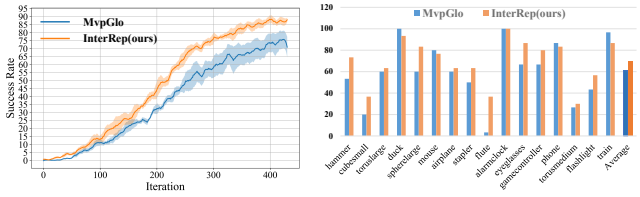


Fig. 6: Success rates (%) of grasping with Allegro. The left is the training results and the right is the testing results.

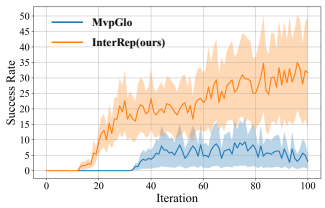


Fig. 7: Success rates (%) of Drawer Opening.

different robotic hands as well as different manipulation tasks.

Objects Grasping with Allegro [1]**.** This experiment aims to demonstrate the adaptability of our method to the robotic hand with different morphology. For grasping with Allegro, the method and simulation environment setup remain largely consistent with the Adroit Hand environment. However, due to the differences in the morphology of robotic hands and in order to expedite training, we simplify the grasping task to train only four objects with different shapes in GRAB [25]: apple, banana, camera, and stamp. The mass is reduced to 0.03 kg, and the horizon is shortened to 60 steps, any trajectory in which the total number of successful steps exceeds 10 is considered as a successful grasp. The objects used for testing are 16 unseen objects from GRAB [25]. Fig. 5 shows the grasping effectiveness and Fig. 6 illustrates the performance of our method compared to **MvpGlo**. The results indicate that our representation method transfers well to grippers of different shapes as it emphasizes the closest distance between the gripper and the object. Despite variations in gripper shapes, the regions closest to the gripper on the object typically remain consistent, reducing sensitivity to gripper shapes.

Drawer Opening. This experiment demonstrates that our proposed representation method also shows some effectiveness in the task of drawer opening in MetaWorld [33], as shown in Fig. 7. Different from the RL method introduced in III-D, we train the manipulation policy from scratch RL, which means the BC stage and demonstrations are excluded from DAPG [20]. The gradient is changed as:

$$g'_{\text{aug}} = \sum_{(s,a)\in\rho_\pi} \nabla_\theta \ln \pi_\theta(a \mid s) A^{\pi_\theta}(s,a). \tag{10}$$

D. Real Robot Experiments

To demonstrate our proposed method functions on a real robot as well, we apply the trained policy of Allegro grasping to the real robotic system, which consists of an Allegro Hand and a Unitree Z1 arm [2]. We test the banana and the camera in the real world. Since it is challenging and costly to train an RL policy in the real world, we assume the object CAD models are known. The results shown in Fig. 8 and Tab. II demonstrate the effectiveness of our proposed representation on real robots.

TABLE II: Success rates (%) of policies tested on the real robotic hand.

Methods	camera	banana	Average
MvpGlo	50	60	55
InterRep(ours)	**70**	**70**	**70**

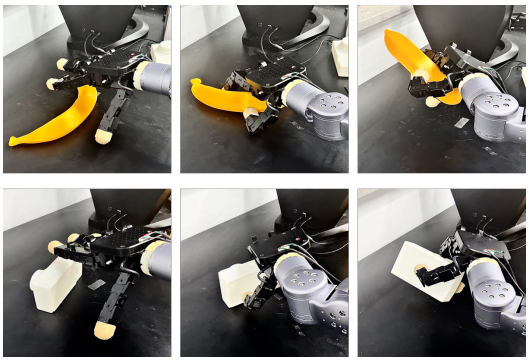


Fig. 8: **Real-world Grasping.** We demonstrate the grasping performance of our method on a real robot.

V. Conclusion

In this paper, we propose a novel image representation for grasping combining pre-trained models with dynamic interaction information. On the one hand, we utilize the encoder of a pre-trained model to extract representations. On the other hand, during each step of reinforcement learning, we extract features from the pixels of the object nearest to the region of the robotic hand, capturing dynamic interaction information as well as the local shape information of the object. This approach not only enhances the sample efficiency of RL training but also improves generalization across objects of different shapes.

Limitations and Future Work. This work primarily validates the effectiveness of our representation method in grasping tasks, which focus on close-range interaction and contact details, typically occurring in scenarios with a limited field of view. However, as concluded in [34], [35], it is evident that pre-training models exhibit variations in performance across diverse tasks. Therefore, we aspire to delve further into discussing the applicability of our proposed representation method across various manipulation tasks or enhance dynamic interaction representations to accommodate a broader range of robot tasks.

[1]https://www.wonikrobotics.com/research-robot-hand

[2]https://www.unitree.com/arm/

REFERENCES

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[2] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[5] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 892–909.

[6] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 359–17 371.

[7] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[8] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Conference on Robot Learning*. PMLR, 2023, pp. 416–426.

[9] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.

[10] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," *arXiv preprint arXiv:2302.12766*, 2023.

[11] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman, "Liv: Language-image representations and rewards for robotic control," *arXiv preprint arXiv:2306.00958*, 2023.

[12] Y. Zhu, A. Joshi, P. Stone, and Y. Zhu, "Viola: Imitation learning for vision-based manipulation with object proposal priors," *arXiv preprint arXiv:2210.11339*, 2022.

[13] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang, "Graph inverse reinforcement learning from diverse videos," in *Conference on Robot Learning*. PMLR, 2023, pp. 55–66.

[14] V. Kumar, Z. Xu, and E. Todorov, "Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands," in *ICRA*. IEEE, 2013, pp. 1512–1519.

[15] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine, "Learning invariant representations for reinforcement learning without reconstruction," *arXiv preprint arXiv:2006.10742*, 2020.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[17] Y. Lu, Z. Zhong, and Y. Shu, "Multi-View Domain Adaptive Object Detection in Surveillance Cameras," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[18] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[19] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, "Imitation learning: A survey of learning methods," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–35, 2017.

[20] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *RSS*, 2018.

[21] Y.-H. Wu, J. Wang, and X. Wang, "Learning generalizable dexterous manipulation from human grasp affordance," in *CoRL*. PMLR, 2023, pp. 618–629.

[22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[23] H. K. Cheng and A. G. Schwing, "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *European Conference on Computer Vision*. Springer, 2022, pp. 640–658.

[24] Q. Liu, Y. Cui, Q. Ye, Z. Sun, H. Li, G. Li, L. Shao, and J. Chen, "Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3153–3160.

[25] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *ECCV*, 2020.

[26] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," in *ICRA*, 2020.

[27] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, Nov. 2017.

[28] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IROS*. IEEE, 2012, pp. 5026–5033.

[29] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, A. Gupta, and D. Fox, "Dextransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations," *arXiv preprint arXiv:2209.14284*, 2022.

[30] W. Wohlkinger, A. Aldoma, R. B. Rusu, and M. Vincze, "3dnet: Large-scale object class recognition from cad models," in *ICRA*, 2012.

[31] X. Wei, M. Liu, Z. Ling, and H. Su, "Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–18, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[33] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Conference on robot learning*. PMLR, 2020, pp. 1094–1100.

[34] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil, *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[35] Y. Hu, R. Wang, L. E. Li, and Y. Gao, "For pre-trained vision models in motor control, not all policy learning methods are created equal," *arXiv preprint arXiv:2304.04591*, 2023.