Proceedings of 2021 IEEE
International Conference on Mechatronics and Automation
August 8 - 11, Takamatsu, Japan

Vision-Based One-Shot Imitation Learning Supplemented with Target Recognition via Meta Learning

Xuyun Yang[1], Yueyan Peng[1], Wei Li[1,2], James Zhiqing Wen[1], Decheng Zhou[1]
[1]Engineering Research Center for Intelligent Robotics, Jihua Laboratory, Guangdong, China
[2]Academy for Engineering and Technology, Fudan University, Shanghai, China
{yangxy & pengyy & liwei & wenzq & zhoudc}@jihualab.com, fd_liwei@fudan.edu.cn

2021 IEEE International Conference on Mechatronics and Automation (ICMA) | 978-1-6654-4101-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICMA52036.2021.9512607

***Abstract*—In this paper, an end-to-end meta imitation learning method supplemented with target recognition (TaR-MIL) is proposed for one-shot learning. This approach divides the procedure of imitating from demonstrations into two parts: distinguishing the target object from distractors and executing the correct actions. Accordingly, the objective of imitation is defined as the combination of target recognition and behavior cloning. Specifically, a target recognition module is adopted in the model architecture, which helps to extract useful information about tasks from observations during training. After training with demonstrations of various tasks via meta learning, a policy capable of solving new tasks given one demonstration is obtained. The real-world experiments on a UR10e robot arm illustrate that, the derived policy manages to perform placing tasks in new scenarios or with new objects after one video demonstration is given, which verify the effectiveness of the proposed method.**

***Index Terms*—Robotics, One-shot imitation learning, Meta learning**

I. Introduction

Recently, intelligent robots are expected to perform complex manipulations in unstructured environments. Usually, it is difficult to manually define a general relationship between environments and robot behavior for a wide range of tasks. Taking insights in human learning from demonstrations and carrying out behavior cloning, *imitation learning* has gained much success and proved to be promising for robots to execute correctly after learning from experts [1]. This learning-based paradigm produces an end-to-end policy which directly maps from perceptual information to action control based on the collected data. In order to generalize over multiple tasks, data augmentation and domain randomization are often used to improve the policy performance of various new tasks [2]–[4].

Further, inspired by the human ability to quickly adapt to novel tasks with previous experience, researchers have proposed some algorithms under the paradigm of *meta learning* with improved learning efficiency which aims to produce a structure with capabilities of fast adaptation. Combining meta learning and imitation learning, meta imitation learning (MIL) [5] offers an effective paradigm for deriving a policy capable of few-shot or one-shot learning, where data efficiency is improved when applying the policy to novel tasks. Considering learning from human, domain adaptive meta learning (DAML) [6] is proposed, where the trained prior policy can adapt to new tasks after learning from one human demonstration.

In this paper, a target recognition-based meta imitation learning (TaR-MIL) approach is proposed. We suppose that when humans try to imitate behavior from others, they will figure out the target object from irrelevant distractors and the corresponding actions to perform the manipulation. Therefore, based on the DAML framework, the proposed TaR-MIL method adopts a target recognition module in the meta-training period, i.e., the objective for post-update is not only defined as the behavior cloning loss (for correct control commands), but also supplemented with the target recognition loss (for identifying the target objects). The target labels for the recognition loss are only needed during training which can help improve the policy to understand the task. This approach can still derive an end-to-end policy for one-shot learning from robot and human.

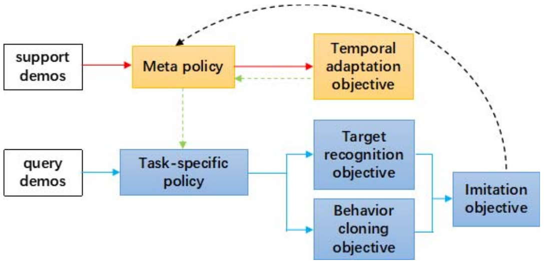

support demos
Meta policy
Temporal adaptation objective
query demos
Task-specific policy
Target recognition objective
Behavior cloning objective
Imitation objective

Fig. 1. The framework of policy training in the proposed TaR-MIL method. The red arrows show the task adaptation procedure using the support demos in the inner loop. The green dash arrows show the gradient for computing the task-specific parameters. In the outer-loop using the query demos (shown as the blue arrows), our method separates the imitation objective into recognizing target and motion cloning, which is used for meta-parameters updating (shown as the black dash arrow). After training, the learned meta policy is capable of solving new tasks after being given one video demonstration and taking gradient steps to obtain the task-specific policy.

A. Related Work

For research on training the robots to perform complex manipulations, there exists extensive works about *imitation learning* [7] for robotic learning from demonstrations, which is more preferable to some traditional model-based approaches, as it is usually difficult to generate a general expression to represent the relationship between the robot configurations and actions in various environments. While the demonstration data can be collected by teleoperation [8], [9], kinesthetic teaching, or measuring device on the instructors [10], recently more researchers have concerned more about learning from videos to find out the correspondence between the visual information and the robot actions.

Specifically, learning from videos of operation by human is one of the substantial issues. Some studies use a hand tracking module to recognize the human activities and then figure out the corresponding robot actions. One limitation of this approach is the reliance on the accuracy of the tracking system. Some works produce an end-to-end policy directly mapping from visual observations to robotic action commands by combining imitation learning and inverse reinforcement learning [11]–[14]. By defining a reward function based on human demonstrations, the policy is optimized under the paradigm of trial and error and finally achieves the desired performance.

However, the policies derived from the methods above possess limited ability to generalize on various scenarios and objects. During training, large amounts of training data should be provided to improve the policy capabilities for multiple tasks. Further, when encountering new tasks, it might be necessary to collect batches of data for learning to acquire the available new mappings.

Considering the model capabilities of generalizing over multiple tasks, some researches utilize multi-task learning approach by using information from multiple tasks to obtain policies available for similar new tasks [15]. Others combine the paradigm of *meta learning* [16]–[18] with imitation learning (e.g., meta imitation learning, MIL), which offer some promising approaches to obtain policies with fast adaptation to new tasks after giving one or a few demonstrations. Furthermore, considering the domain shift between human demonstrations and robot executions, Yu et al. proposed the domain adaptive meta learning (DAML) approach by extending MIL with a temporal adaptation loss.

Inspired by finding the available representations of the tasks from demonstrations, a meta imitation learning method supplemented with target recognition (TaR-MIL) is proposed in this paper. The proposed method adopts a target recognition module for post-update during the meta-training period, in order to improve the policy to extract useful feature information for the task. In the inner loop of meta-training, task adaptation is first performed using demonstration videos. Then in the outer loop, the policy tries to figure out the target object from other distractors in the scenarios and produce the correct actions concurrently.

B. Contribution and Organization

The main contributions of the paper are as follows:

- A one-shot visual imitation learning method supplemented with target recognition is proposed using a framework of meta learning. The method is capable of inferring an end-to-end policy which directly maps the visual observations to robot control commands, after using one demonstration of robot or human to update the trained prior policy.
- The real-world experiments are carried out, which use a robot arm to perform placing tasks. The experimental results substantiate the effectiveness of our method, illustrating that the derived policy is capable of learning from one demonstration video of robot or human to solve novel tasks. The comparison results show that our method outperforms the naive MIL method, verifying the importance of the supplemented target recognition module of our method.

The remainder of this paper is organized as follows. In Section II, backgrounds about meta learning algorithm (model agnostic meta learning, MAML) and imitation learning via meta learning (MIL and DAML) are presented. In Section III, our TaR-MIL method is proposed. In Section IV, real-world experiments of one-shot learning from robot and human for placing tasks are presented, which verify the effectiveness and robustness of our method. In Section V, conclusions of this paper are given.

II. Background

In this section, a gradient-based meta-learning method, the model-agnostic meta learning algorithm (MAML) and its extension to imitation learning, meta imitation learning (MIL) [5] and domain adaptive meta learning (DAML) [6] are presented. Meta learning aims to handle with new tasks which have not been seen before during meta-testing. MAML attempts find a prior model with parameters $\boldsymbol{\theta}$ which can quickly adapt to new tasks. Given some demonstrations of the new tasks, the task-specific parameter $\boldsymbol{\phi}$ is obtained after taking steps of gradient descent on $\boldsymbol{\theta}$, and the model with $\boldsymbol{\phi}$ is available to handle new objects.

In MAML, a number of tasks are involved, and a distribution $p(\mathcal{T})$ over tasks is assumed. During meta-training, a set of tasks $\{\mathcal{T}_i\}$ are drawn from $p(\mathcal{T})$, with support set $D^{tr}_{\mathcal{T}_i}$ and query set $D^{val}_{\mathcal{T}_i}$ prepared for each task. The support set and the query set are used for two phases, respectively. In the inner loop, the model f is initialized with meta-parameter $\boldsymbol{\theta}$. For each sampled task $\mathcal{T}_i$, the adaptive parameter $\boldsymbol{\phi}_i$ is computed by $\boldsymbol{\phi}_i = \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathcal{T}_i}(f_{\boldsymbol{\theta}}, D^{tr}_{\mathcal{T}_i})$, where α is the stepsize of the gradient descent. In the outer loop, the task-specific parameter $\boldsymbol{\phi}_i$ is evaluated on the query set for each sampled task $\mathcal{T}_i$, and the performance across tasks are used to optimize the meta-parameter, i.e., the objective of outer loop is:

$$\min_{\boldsymbol{\theta}} \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}_{\mathcal{T}_i}(f_{\boldsymbol{\phi}_i}, D^{val}_{\mathcal{T}_i}), \tag{1}$$

After combining the objective in the inner loop and the outer loop, the optimization objective of the model parameters is expressed as

$$\min_{\boldsymbol{\theta}} \sum_{\mathcal{T}} \mathcal{L}(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \mathcal{D}^{tr}_{\mathcal{T}}), \mathcal{D}^{val}_{\mathcal{T}}) \tag{2}$$

where $\mathcal{L}$ is defined by supervised learning loss for behavior cloning, e.g., mean square error.

MIL extends MAML to imitation learning, with a two-head model architecture for learning from visual demonstrations. In the training period of MIL, one head is used for action output in the inner loop and the other is used in the outer loop. Considering learning from human videos which only provide sequence of images, the DAML approach is proposed. In DAML, temporal convolutional module is constructed for the designated temporal adaptation objectives $\mathcal{L}_{\psi}$ (ψ is the parameter of the temporal module) as the loss of inner loop. Thus, the corresponding objective of the meta-parameter $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$ is defined as follows:

$$\min_{\boldsymbol{\theta},\boldsymbol{\psi}} \sum_{\mathcal{T}} \mathcal{L}(\boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, \mathcal{D}^{tr}_{\mathcal{T}}), \mathcal{D}^{val}_{\mathcal{T}}) \tag{3}$$

Although this method manages to learn from one robot or human demonstration to solve new tasks, it relies much on the temporal loss to provide behavior information extracted from a video. During the outer loop in the training period, DAML only focuses on cloning the demonstrated robot action, while our method also tries to distinguish the corresponding target object among multiple distractors in the task. The supplemental loss of target recognition can help the policy to understand the behavior as the demonstrator shows in the video.

III. Methodology

In this section, we first describe the problem of one-shot imitation learning. We suppose that the proposed method manages to infer the policy available for novel situations after one video of robot or human demonstration is given. Then, we introduce the proposed meta imitation learning method with supplemented target recognition (TaR-MIL) and present the corresponding three-head network architecture for the policy.

A. Problem Description

In this paper, the problem of learning from videos of robot or human is studied. We consider this problem as learning to deduce the policy for new tasks involving new objects in dynamic environments, after one demonstration is given.

The training dataset $D = \{D^{tr}, D^{val}\}$ containing support set D^{tr} and query set D^{val} is provided. Assuming that there is a distribution $p(\mathcal{T})$ over the tasks $\mathcal{T}$, a set of tasks $\{\mathcal{T}_i\}$ is sampled from $p(\mathcal{T})$. For each task $\mathcal{T}_i$, a support set $D^{tr}_{\mathcal{T}_i}$ and a query set $D^{val}_{\mathcal{T}_i}$ is prepared. Within the support set $D^{tr}_{\mathcal{T}_i}$, videos of multiple demonstrations are contained, which are sequences of visual observations with length T, i.e., $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$. Within the query set $D^{val}_{\mathcal{T}_i}$, the robot demonstrations contain sequences of image observations and robot actions with length $\hat{T}$, i.e., $\{\boldsymbol{o}_1, \boldsymbol{a}_1, \ldots, \boldsymbol{o}_{\hat{T}}, \boldsymbol{a}_{\hat{T}}\}$.

By using the support set D^{tr} and query set D^{val} of different tasks, a meta-policy is obtained via the training paradigm of meta learning. When deploying the learned meta-policy, the policy is first adopted to the given demonstration of a new task and it is updated based on the adaptation objective. Afterwards, the updated task-specific policy is capable of solving the new task.

B. Target Recognition-based Meta Imitation Learning

We define a new task as a robot encounters new target objects or new scenarios. For one-shot imitation learning, the robot is supposed to be capable of solving new tasks after being provided with one demonstration video of new tasks.

Algorithm 1: One-Shot Meta Imitation Learning with Target-Recognition

```
Input: Sets of tasks $\{\mathcal{T}_i\}$, with $\mathcal{T}_i \sim p(\mathcal{T})$;
       demonstrations for support sets and query sets;
       stepsize $\alpha$ and $\beta$.
Output: Prior policy with the trained meta-parameters
        $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$
// Training procedure.
Initial meta parameters $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$ randomly;
while iteration $it = 1, 2, \ldots, I$ do
    // Outer-loop
    Sample batch of meta-training tasks $\mathcal{T}_i \sim p(\mathcal{T})$;
    foreach task in $\mathcal{T}_i$ do
        // Inner-loop
        Sample support set $D^{tr}_{\mathcal{T}_i}$;
        Evaluate task adaptive objective: $\mathcal{L}_{\boldsymbol{\psi}}(f_\theta, D^{tr}_{\mathcal{T}})$
          Update task-specific parameters:
          $\boldsymbol{\phi} = \boldsymbol{\theta} - \alpha\nabla L_{\boldsymbol{\psi}}(\boldsymbol{\theta}, D^{tr}_{\mathcal{T}})$;
        Sample query set $D^{val}_{\mathcal{T}_i}$;
        Evaluate task performance objective:
          $\mathcal{L}_{val}(\boldsymbol{\phi}, D^{val}_{\mathcal{T}_i}) =$
          $\mathcal{L}_{TaR}(\boldsymbol{\phi}, D^{val}_{\mathcal{T}_i}) + \mathcal{L}_{BC}(\boldsymbol{\phi}, D^{val}_{\mathcal{T}_i})$;
    end
    Update meta-parameters by:
      $\{\boldsymbol{\theta}, \boldsymbol{\psi}\} \leftarrow \{\boldsymbol{\theta}, \boldsymbol{\psi}\} - \beta\nabla_{\boldsymbol{\theta},\boldsymbol{\psi}} \sum_{\mathcal{T}_i} \mathcal{L}_{val}(\boldsymbol{\phi}, D^{val}_{\mathcal{T}_i})$;
end
// Testing procedure.
Prepare sequence of observations of demonstration
  $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ for the new task $\mathcal{T}_{test}$ for meta-testing;
Obtain the updated task-specific parameters
  $\boldsymbol{\phi} = \boldsymbol{\theta} - \alpha\nabla\mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, D^{tr}_{\mathcal{T}_{test}})$;
Given sequence of observation $\{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_{\hat{T}}\}$, output
  the actions from the inferred control policy
  $\pi(\boldsymbol{a}_i|\boldsymbol{o}_i, \boldsymbol{\phi})$.
```

Generally in a scenario, in addition to the targets, there are other distractors. When humans learn from a video, they recognize the target (what) and then generate the corresponding behavior (how). Inspired by human learning, the proposed method adopts this by considering a target recognition-based loss during the evaluation of the updated parameters, since recognizing the target object is important for the robot to reproduce the corresponding behavior in dynamic environments. Based on the training phase of the meta imitation learning method [5], we define the imitation loss for post-update in the outer-loop of training as:

$$\mathcal{L}_{\text{val}}(\boldsymbol{\phi}, D^{val}) = \mathcal{L}_{\text{TaR}}(\boldsymbol{\phi}, D^{val}) + \mathcal{L}_{\text{BC}}(\boldsymbol{\phi}, D^{val}) \tag{4}$$

where $\mathcal{L}_{\text{TaR}}$ is the objective of target recognition aiming to guide the policy to figure out the traget objects, and $\mathcal{L}_{\text{BC}}$ is the objective of behavior cloning aiming to lead the policy to produce the correct actions. With $\mathcal{L}_{\text{val}}$, gradient steps are taken with respect to the initial meta-parameters $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$. Since the target recognition objective is defined as a supervised loss function with the demo label, the designated imitation objective above partly provides more guidance for learning to correctly perform the tasks. For pre-update in the inner loop of

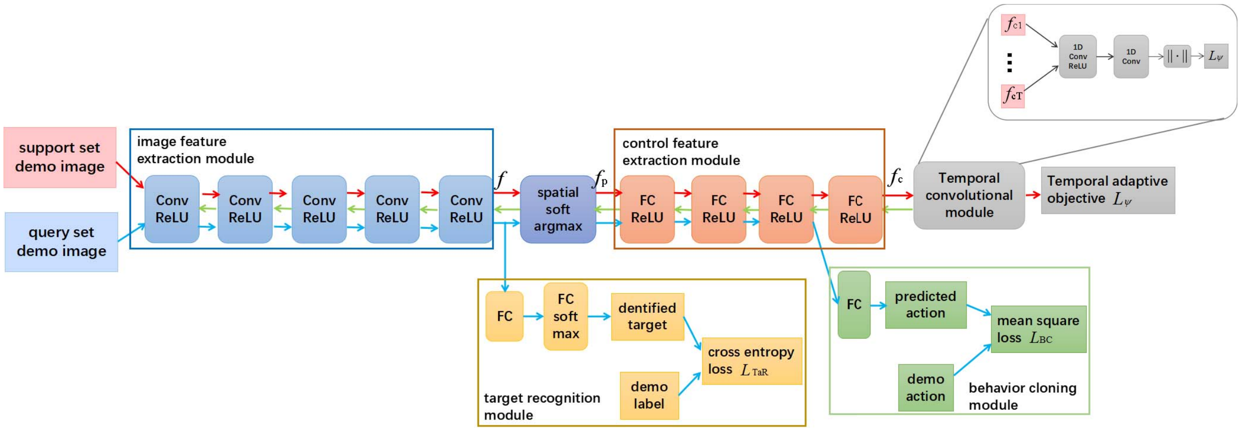


Fig. 2. Network architecture of the policy. The red arrows denote the forward calculation of the meta-policy in the inner loop. The green arrows denote the backward calculation for gradient in the inner loop to obtain the updated task-specific policy. The blue arrows denote the forward calculation of the task-specific policy in the outer loop.

training, our method retain the temporal convolutional module [19] for adaption loss in DAML [6] which is denoted as $\mathcal{L}_\psi$. In summary, the objective of the whole meta-trainig period of our method is formulated as:

$$\min_{\boldsymbol{\theta},\boldsymbol{\psi}} \sum_{\mathcal{T}} \mathcal{L}_{\text{val}}(\boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\theta}, D_{\mathcal{T}}^{tr}), D_{\mathcal{T}}^{val}) \tag{5}$$

The procedure of optimizing this objective function and deriving the meta-policy is summarized in Algorithm 1. Demonstration data for multiple tasks including support set and query set are prepared for meta-training. In the inner loop for pre-update, the task-specific policy with parameters $\boldsymbol{\phi}$ are inferred from $\boldsymbol{\theta}$ by using the temporal adaptation loss $\mathcal{L}_\psi(\boldsymbol{\theta})$. In the outer loop for post-update, the meta parameters $\boldsymbol{\theta}$ is updated based on the validation error $\mathcal{L}_{val}(\boldsymbol{\phi})$.

In the proposed TaR-MIL method, the network architecture of the policy is shown in Figure 2. During the inner loop, frames of image observations from human demonstrations of tasks concurrently go through the feature extraction module and the final adaptive loss module, and then the task adaptive loss $\mathcal{L}_\psi(\phi)$ is obtained ($\boldsymbol{\phi}$ is equal to θ initially). After taking one or multiple gradient steps based on the minimum target $\mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\phi})$, the task-specific model parameter can be derived. During the outer loop, the current image observation is taken as the input of the model with the updated parameters $\boldsymbol{\phi}$, which outputs the policy $\pi(\boldsymbol{a}|\boldsymbol{o})$. Detailed introduction of the network architecture will be presented in the next section.

C. Network Architecture

Compared with the previous two-head structure, a three-head network architecture of our method is presented in Figure 2. One head denoted as a temporal convolutional module is for computing the adaptive loss of the video demonstrations to obtain the task-specific parameters in the inner loop of training. The other two heads, denoted as the target recognition module and the behavior cloning module, compute the meta-objective containing the objective of recognizing target and the objective of behavior cloning, respectively.

The image observations go through the former image feature extraction module and the latter modules including the pre-update adaptation module, the target recognition module and the control feature extraction module. In the image feature extraction module, four 2D convolutional layers with ReLU (denoted as Conv-ReLU) are adopted and the image feature $\boldsymbol{f}$ is obtained. Followed by the spatial soft-argmax computation, the position feature vector $\boldsymbol{f}_\text{p}$ is obtained as the output of the image feature extraction module. The control module takes $\boldsymbol{f}_\text{p}$ as the input and generates control feature f_c by three fully-connected layers with ReLU (denoted as FC-ReLU). As the effectiveness of the temporal convolutions for learning from demonstration video without actions is verified in Yu et al. work [6], we retain the temporal loss for the task adaption objective. During the inner loop of meta-training, batches of $\boldsymbol{f}_\text{c}$ of sequence of images from demonstrations go through the temporal convolutional layers concurrently. During the outer loop, $\boldsymbol{f}_\text{c}$ for each image observation $\boldsymbol{o}_t$ goes through a fully-connected layer which outputs the robot control policy $\pi(\boldsymbol{a}_t|\boldsymbol{o}_t)$, and the behavior cloning loss can be computed. On the other hand, the raw feature $\boldsymbol{f}$ goes through the target recognition module adopting two fully-connected layers, which predicts the target object in a scenario for computing the target recognition loss.

Note that, though the label of the target object in the demonstration should be provided for the target recognition objective during meta-training, we do not need the object label during meta-testing. Thus, except the demo video of image sequence, no more task-relative information is provided for the prior model to adapt to the new task, which retain the flexibility of the proposed method.

IV. Experiments

In this section, real-world experiments are conducted with a 6-DoF UR10e robot. Considering that garbage sorting is an important issue in our life, we perform our method on a placing imitation task where the demonstrator teaches the robot how to sort garbage. In this task, the robot learns to

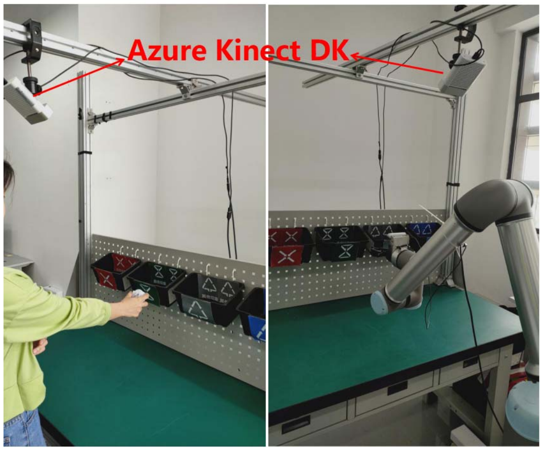


Fig. 3. Experimental platform for the placing task, which is used for collecting human demos (left) and robot demos (right). Policy evaluation is conducted on this platform.

approach the corresponding containers as the demonstrator arrives in the video. The experimental results substantiate the effectiveness of the proposed TaR-MIL method. Our method manages to handle dynamic scenarios or new containers which have not been seen during training, after giving one-shot demonstration of a video from robot or human. Besides, comparison experiments with the *naive MIL* method are also conducted, which abbreviates the target recognition module in Figure 2 and only uses the behavior cloning loss for the post-update objective.

A. Training Configurations

We use the Azure Kinect camera to collect image observations of demonstration of support set and query set for the meta-training period. Concurrently, we use UR10e robot to collect the robot action data. The experimental platform for collecting demos and evaluation after training are shown in Figure 3. Four boxes for different kinds of garbage are considered during training, including recyclable, harmful, food and other waste. We define a task as arriving at the box as demonstrated to put objects into it, and the other boxes positioned randomly in the scenarios are regarded as the distractors in this task. Besides, we collect demonstration videos of sequences of images with length $T = 10$. For imitation, we only sample the first $\hat{T} = 10$ image observations $\boldsymbol{o}_t$ $(1 \leq t \leq \hat{T})$ with action $\boldsymbol{a}_t$. We consider discrete action control for robot, i.e., $\boldsymbol{a}_t$ is defined as a direction vector indicating that the robot should take to arrive at the target box (e.g., when the robot should arrive at the furthest right box, $\boldsymbol{a}_t = [0, 0, 0, 1]$). For target cognition, we also define each target object with a vector, i.e., $\boldsymbol{c}_{\mathrm{r}} = [1, 0, 0, 0]$ for box of recyclable waste, $\boldsymbol{c}_{\mathrm{f}} = [0, 1, 0, 0]$ for food waste, $\boldsymbol{c}_{\mathrm{h}} = [0, 0, 1, 0]$ for harmful waste, and $\boldsymbol{c}_{\mathrm{o}} = [0, 0, 0, 1]$ for other waste.

Since using UR10e robot for data collection is time-consuming, we first use the robot arm to collect a primary small dataset of 288 demonstrations. Afterwards, we augmented the collected images to extend our dataset, including adding random border, setting random brightness factor and adding random Gaussian noises. Concurrently, we collect 96 human demonstrations, which are augmented in a similar way to the robot dataset augmentation.

The detailed training settings are as follows:

- Image settings: the collected RGB images are of size 256×144.
- Network settings: (a) image feature extraction module: 4 convolutional layers with 64 3×3 filters and stride 3, followed by 1 convolutional layer with 64 3×3 filters and stride 1; (b) control feature extraction module: 4 fully connected layers of size 50; (c) temporal convolutional module: 1 layer of 10 10×1 filters, followed by 1 layer of 10 1×1 filters; (d) target recognition module: 1 fully-connected layer of size 64 followed by 1 fully-connected layer of size 4; (e) behavior cloning module which outputs control order: 1 fully-connected layer of size 4.
- Hyperparameters: stepsize for inner loop $\alpha = 0.005$ and outer loop $\beta = 0.001$; gradient clipping range of $[-10, 10]$; meta batch size of 4; one demonstration video and one different evaluation video for each sampled task; training iteration $I = 75k$.

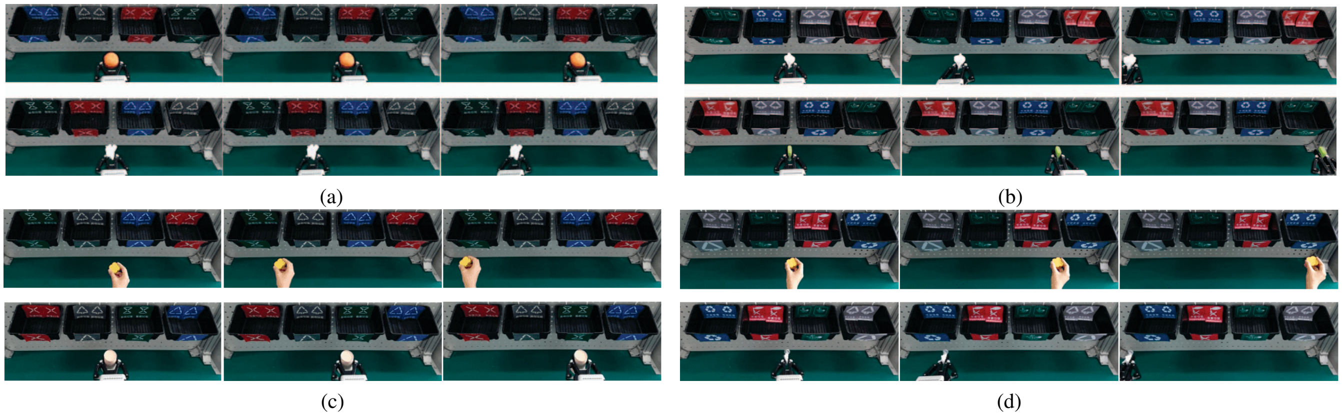
B. Experimental Results

Evaluation is carried out under two circumstances: (a) Using boxes with textures which have been seen during meta-training, but conducted in dynamic scenarios, e.g., variable illumination, slight change of perspective. (b) Using boxes with different patterns which have not been seen during training. As shown in Figure 4, the trained meta-policy is capable of inferring a task-specific model to solve new scenarios or new containers (i.e., with patterns which have not been seen during training), after one demonstration video of human or robot is provided. We conducted 2304 trials for new scenarios and 64 trials for new containers. With robot demonstrations, the robot can arrive at the corresponding boxes correctly with 93% success rate and 83% success rate for new scenarios and new boxes, respectively. With human demonstrations, the robot can perform 87% and 77% success rate for new scenarios and new boxes, respectively. These results verify the effectiveness of our TaR-MIL method.

Furthermore, comparison results with the naive MIL method which abbreviates target recognition are presented in Table I. It is shown the performance of the naive MIL method is worse than our method. This verifies the importance and validity of seperating imitation procedure of recognizing target boxes and executing the respective actions.

V. Conclusions

This paper proposed an approach for one-shot imitation learning which enables the robot to handle new situations by giving one demonstration video. We utilize the framework of meta imitation learning with prior knowledge over different tasks. Our approach introduces a target recognition module for figuring out the target objects from other distractors


(a)

(b)

(c)

(d)

Fig. 4. Meta-policy learns from robot and human during evalution. The top rows are figures of the provided demos, and the bottom rows are the robot imitations. The robot manages to arrive at the corresponding containers as the demonstrators perform in videos. Note that, the objects held by the human or the robot can be different, which are not related to the approaching task. (a) Learning from a robot video in a new scenario. (b) Learning from a robot video with new containers. (c) Learning from a human video in a new scenario. (d) Learning from a human video with new containers.

TABLE I
COMPARISON TESTING RESULTS.

method	TaR-MIL (Ours)	naive MIL
new scenarios (learning from robot)	**93**%	25%
new containers (learning from robot)	**83**%	25%
new scenarios (learning from human)	**87**%	25%
new containers (learning from human)	**77**%	25%

during post-update in the training period, which improves the imitation performance. The real-world placing experiments using UR10e robot arm verify the effectiveness of our method and demonstrate the importance of the supplemented target recognition module. In the future, the proposed method can be further studied about the data efficiency issue during the training process.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge the funding support from Ji Hua Laboratory through the Self-Programming Intelligent Robot Project (No. X190101TB190).

REFERENCES

[1] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne, “Imitation learning: A survey of learning methods,” *ACM Computing Surveys*, vol. 50, no. 2, 2017.

[2] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, “Reinforcement learning with augmented data,” *ArXiv*, vol. abs/2004.14990, 2020.

[3] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, “Visual imitation made easy,” *ArXiv*, vol. abs/2008.04899, 2020.

[4] I. Kostrikov, D. Yarats, and R. Fergus, “Image augmentation is all you need: Regularizing deep reinforcement learning from pixels,” *ArXiv*, vol. abs/2004.13649, 2020.

[5] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, “One-shot visual imitation learning via meta-learning,” in *Conference on Robot Learning*, 2017, pp. 357–368.

[6] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine, “One-shot imitation from observing humans via domain-adaptive meta-learning,” *ArXiv*, vol. abs/1802.01557, 2018.

[7] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, “Survey of imitation learning for robotic manipulation,” *International Journal of Intelligent Robotics and Applications*, vol. 3, pp. 362 – 369, 2019.

[8] T. Zhang, Z. McCarthy, O. Jow, D. Lee, K. Goldberg, and P. Abbeel, “Deep imitation learning for complex manipulation tasks from virtual reality teleoperation,” *IEEE International Conference on Robotics and Automation*, pp. 1–8, 2018.

[9] A. Mandlekar, D. Xu, R. Martín-Martín, Y. Zhu, L. Fei-Fei, and S. Savarese, “Human-in-the-loop imitation learning using remote teleoperation,” *ArXiv*, vol. abs/2012.06733, 2020.

[10] M. Edmonds, F. Gao, X. Xie, H. Liu, S. Qi, Y. Zhu, B. Rothrock, and S. Zhu, “Feeling the force: Integrating force and pose for fluent discovery through imitation learning to open medicine bottles,” *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3530–3537, 2017.

[11] J. Hua, L. Zeng, G. Li, and Z. Ju, “Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning,” *Sensors*, vol. 21, 2021.

[12] R. A. M. Strudel, A. Pashevich, I. Kalevatykh, I. Laptev, J. Sivic, and C. Schmid, “Combining learned skills and reinforcement learning for robotic manipulations,” *ArXiv*, vol. abs/1908.00722, 2019.

[13] M. Mohammedalamen, W. D. Khamies, and B. Rosman, “Transfer learning for prosthetics using imitation learning,” *ArXiv*, vol. abs/1901.04772, 2019.

[14] A. Rajeswaran, V. Kumar, A. Gupta, J. Schulman, E. Todorov, and S. Levine, “Learning complex dexterous manipulation with deep reinforcement learning and demonstrations,” *ArXiv*, vol. abs/1709.10087, 2018.

[15] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, “Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration,” *IEEE International Conference on Robotics and Automation*, pp. 3758–3765, 2018.

[16] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning*, 2017, pp. 1126–1135.

[17] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” *ArXiv*, vol. abs/2004.05439, 2020.

[18] A. Antoniou, H. Edwards, and A. Storkey, “How to train your maml,” *ArXiv*, vol. abs/1810.09502, 2019.

[19] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” in *ISCA Speech Synthesis Workshop*, 2016.