IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

Few-Shot Instance Grasping of Novel Objects in Clutter

Weikun Guo, Wei Li, Ziye Hu, Zhongxue Gan

***Abstract*—Instance grasping, which aims to grasp a specific object out of clutter, is a fundamental task within robotics. However, allowing a robot to quickly learn to perform instance grasping for new, previously unseen objects remains challenging. In this work, we present an instance grasping meta-learning framework (IGML), a simple yet effective end-to-end approach that not only teaches robots to identify novel objects but also how to grasp them. Given only a few examples to specify the grasping point of the target object, our IGML can quickly learn to recognize the target object and grasp it at the demonstrated grasping point by leveraging prior experience. Experimental results on the test sets show that IGML achieves decent success rates in cluttered environments, significantly surpassing state-of-the-art methods. Then we deployed IGML on a UR5 robot arm to handle pick-and-place scenarios and achieved a precision rate of 93.4% and a recall rate of 87.1%.**

***Index Terms*—Perception for grasping and manipulation, deep learning for visual perception, deep learning in grasping and manipulation.**

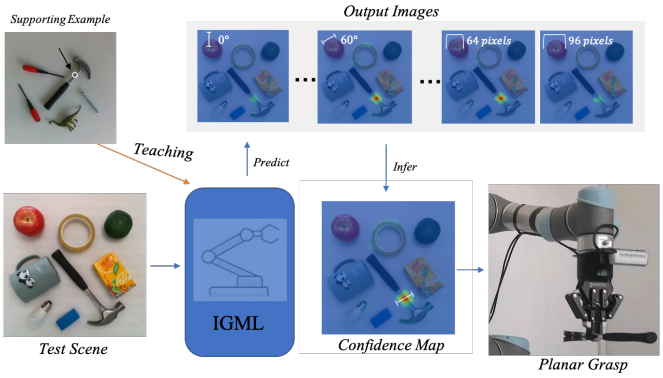

Fig. 1. Given a few examples to specify a grasping point of a novel object, our robot can quickly learn to grasp it out of clutter at the specified grasping point with proper gripper orientation and opening width. Brightness indicates the confidence score.

I. INTRODUCTION

COMPARED to indiscriminate grasping, robotic instance grasping, which aims to grasp a specific object out of a set of distinct objects, is a more challenging task as robots need to localize the target object and infer a grasp pose for the target object. Classical solution for robotic instance grasping requires a recognition module, such as semantic segmentation [1] or object detection [2], to localize the target object, followed by a grasp prediction module to predict the grasp pose. While these methods achieve instance grasping for trained objects, they can not generalize to previously unseen objects. For this reason, Constraint Co-Attention Network (CCAN) [3] and Attribute-Based Robotic Grasping (ABRG) [4] were proposed to tackle the emerging problem of unseen object instance grasping. However, the performance of these methods is usually limited even in non-cluttered environments, mainly due to the accumulation error of recognizing novel objects and grasping.

While significant progress has been made in vision-based robotic grasping, choosing a grasp based on the appearance of objects often fails in some cases for physical or task constraints. For example, robots might spoil objects (e.g., grasping feathers of badminton or body of bulbs), and objects with complex shapes or uneven mass distribution would slip out of the gripper (e.g., grasping the tail of a hammer). Such situations require expert knowledge to teach robots to grasp a suitable portion of the object. Unlike prior works that consider recognition and grasping separately [3], [4], we argue that the success of recognition should promote the success of grasping. To this end, we propose to teach the robot not only to recognize novel objects but also to recognize the target grasping point of the novel objects. As a result, robots tend to perform a good grasp if they localize the target grasping point with the desired accuracy.

Inspired by the fact that humans could fine-tune and re-combine a set of pre-existing skills to learn a new task rather than learning from scratch [5], we propose to achieve instance grasping via meta learning [6], [7], enabling a robot to reuse past experience and, as a result, learn to recognize and grasp novel objects in cluttered environments quickly. Concretely, we trained a meta-policy that can perform 4-DoF planar grasping, enabling the robot to indiscriminately grasp objects in the workspace. As shown in Fig. 1, given one or a few examples to specify the grasping point of the target object, the meta policy can fast adapt (e.g., fine-tuning with one gradient step) to recognize the demonstrated grasping point

Manuscript received: December 29, 2021; Revised March 30, 2022; Accepted April 25, 2022. This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments and was accepted as an oral presentation by ICRA 2023.

This work was supported in part by Ji Hua Laboratory under Projects ID X190021TB190, in part by Shanghai Municipal Science and Technology under Major Project 2021SHZDZX0103, in part by the Science and Technology Commission of Shanghai Municipality under Project 19511132000, in part by the Shanghai Engineering Research Center of AI and Robotics, and in part by the Engineering Research Center of AI and Robotics, Ministry of Education, China.

Corresponding author: Wei Li (email:fd_liwei@fudan.edu.cn). Weikun Guo, Wei Li, Ziye Hu, Zhongxue Gan are with the Academy for Engineering and Technology, Fudan University, Shanghai, China, and also with the Department of Engineering Research Center for Intelligent Robotics, Ji Hua Laboratory, Guangdong, China.

Digital Object Identifier 10.1109/LRA.2022.3174648.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2023, London, UK. Cite as RA-L paper.

of the target object in new environments without forgetting its original grasping ability. We also demonstrated that IGML could be improved to fast adapt to recognize more than one novel instance to increase robots' picking speed.

During fast adaptation, the meta-policy learns from the supporting image (e.g., Fig. 1) where both positive samples (e.g., target objects) and negative samples (e.g., distractor objects) are in the scene. We argue that the meta-policy can either learn from only positive samples or from both positive and negative samples to achieve the instance grasping. The intuition is that learning from both positive and negative samples would help distinguish the target object from distractor objects that appear in the supporting image, while learning from only positive samples can focus on the features of the target object and trivialize other irrelevant features. Considering the loss for fast adaptation would affect the learning from supporting examples, we studied the adaptive loss to allow the meta-policy to learn from positive or negative samples. As a result, the meta-policy can learn efficiently from the supporting examples to recognize a novel object through learning the features of the target object in the supporting images while trivializing other irrelevant features.

In this paper, our main contribution could be summarized as follows:

- A novel and practical end-to-end IGML framework is proposed to enable robots to rapidly adapt to grasp specified novel objects out of the clutter. Compared to previous approaches, the proposed method first allows the robot to perform instance grasping for novel objects in cluttered scenes robustly.
- Instead of only learning to recognize novel objects, IGML also learns to locate specified grasping points from examples. We demonstrated that teaching the robot to grasp at designated grasping points will significantly contribute to grasping success, especially when objects have complex shapes.
- We enrich IGML framework to enable the robot to learn from positive or negative and investigate their advantages in different scenarios. Comprehensive experiments on test sets and the physical world demonstrate the effectiveness and robustness of IGML, which significantly outperforms state-of-the-art methods.

II. RELATED WORK

With the development of vision-based deep learning, large advancements have recently been seen in grasping unknown objects. Broadly, previous works for robotic grasping can be divided into indiscriminate grasping [8]–[13] and instance grasping [1], [3], [4], [14], [15]. Indiscriminate grasping aims to grasp any objects from the workspace indiscriminately. Lenz et al. [8] presented a two-stage method that first generated a set of oriented rectangles as grasping candidates and then ranked these candidates individually to select the best grasp. Kumra et al. [11] proposed an end-to-end model that outputs images representing grasp quality score, gripper orientation, and gripper width, from which they can infer a gripper pose and a confidence score for each pixel. However, the robot does not perform well when grasping complex objects with new geometries, as we have observed in experiments that it is difficult for the model to consistently select a good grasping point from the grasp quality map.

In contrast, instance grasping aims to grasp a specific object or object class out of a set of distinct objects. Jang et al. [14] learned to grasp a specific object class with an end-to-end framework of object detection, classification, and grasp planning. However, the method cannot be generalized to recognize and grasp new object classes. Building on the indiscriminate grasping method [16], Cai et al. [3] proposed CCAN, an attention-based Siamese network that can localize target objects and determine whether a horizontal grasp can hold them. While the robot can localize a novel target object with a rough mask according to a single image, the lack of accurate grasping points leads to a considerable drop in success rate. Yang et al. [4] learned an attribute-based multimodel framework that takes in a text description of the color and shape of the target object to achieve unseen object instance grasping. However, they only demonstrated the effectiveness of their approach on objects with simple shapes and colors (e.g., a red sphere).

Some works are devoted to enabling robots to learn grasping preference [17]–[19]. Helenon et al. [18] trained a model for each object to segment the prohibited and authorized portion of the object. While they can designate a suitable grasping portion of the object to robots, the gripper pose can't be obtained directly by the model, and the model lack recognition ability. In comparison, IGML can fast adapt to recognize and grasp the target object on the designated grasping point with the desired accuracy.

Regarding pick-and-place, many solutions adopt semantic segmentation to detect objects' location and identity, parallel with predicting the grasp pose for the vacuum gripper [20], or followed by model-based grasp planning [21]. While these methods are very efficient for scenarios where objects are all known, they tend to fall short when dealing with applications that constantly encounter new objects, as collecting large amounts of data and re-training models from scratch is very expensive. Instead, Zeng et al. [12] proposed a grasp-first-then-recognize method, taking advantage of the generalization of robotic indiscriminate grasping to grasp an object out of clutter, after which the robot takes a photo of the object in another camera to recognize it by image matching. While this approach is very practical for robotic pick-and-place of novel objects, it shows great limitations when only one or a few objects are needed to be picked out by the robot from a bunch of objects. Berscheid et al. [22] learned pick-and-place objects with precise placing accuracy according to a single, demonstrated goal state. However, it shows limited performance in recognizing novel objects, achieving only a 60% average success rate in 1-out-of-5 selection tasks.

Some approaches concerning meta-learning investigated manipulations of novel objects [23]–[25]. Finn et al. [23] combined meta-learning with imitation to enable robots to learn to push a novel object or place a held object within a novel big bowl from a video demonstration. However, they can only manipulate the target object with a rough location without

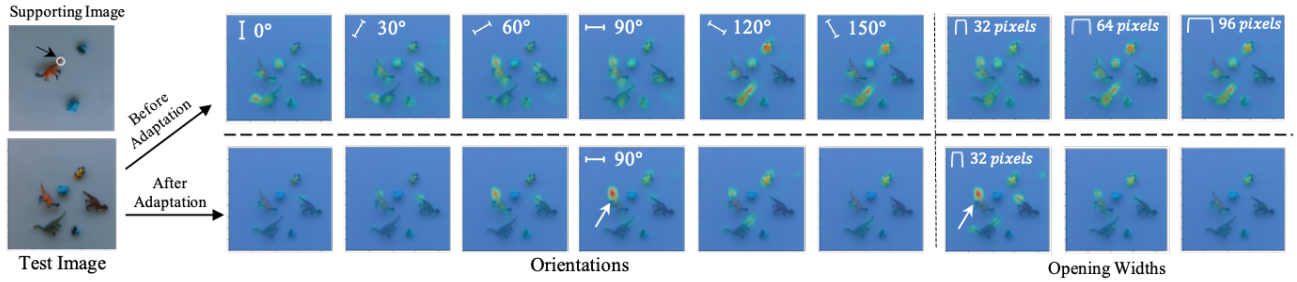

Fig. 2. An example shows the output of IGML before and after adaptation. The adapted policy allows the robot to grasp the target objects at the specified portion. The brightness of the output heat maps represents the confidence score, from which we can infer the instance grasp, e.g., we select the gripper rotation angle of 90° and the smallest gripper opening width (32 pixels) to grasp the tail of the orange dinosaur.

an accurate, specific point of the target object or even gripper pose (e.g., gripper orientation and opening width). Gualtieri et al. [26] proposed a method that learned from simulation to grasp bottles or mugs in clutter and place them upright. However, the placements cannot generalize to novel object classes.

III. METHODOLOGY

A. Problem Overview and Definition

Our goal is to learn a meta-policy f_θ that can quickly adapt to new instance grasping tasks from a few supporting examples. A new instance grasping task means recognizing two novel objects orderly (e.g., o_1 and o_2) and grasping them out of a set of distinct objects. As we can predict two instances, we define the first object as o_1 and the second object as o_2. For simplicity, we define the distractor objects in the adaptation images (supporting examples) as d_{sup}. Much like metric-based meta-learning, we consider the key idea in recognizing novel objects is training the meta-policy for adaptation over various tasks, such that the policy f_θ would learn an internal representation that is broadly suitable for most tasks [7], which we call meta-features, encapsulating object profile, color, and other visual discriminative properties. During fast adaptation, the meta-policy combines and fine-tunes these meta-features to generate the adapted-policy f_ϕ for each task, which has a high output response to the target grasping point of the target object and low output response to distractor objects. Then we fused recognition with grasping through defining grasp primitives, which differ in the gripper orientation or opening width. Fig. 2 shows the nine grasping primitives we defined.

We formally define the task to grasp a novel instance in a top-down manner in the robot coordinate as:

$$IG_r = (L_r, R_r, W_r, S) \tag{1}$$

where $L_r = (x_r, y_r, z_r)$ is the reaching position of the center of the gripper's tip, R_r is the gripper rotation around the z-axis, and W_r is the required opening width for the two-finger gripper. They constitute the end-effector pose $A_r = (L_r, R_r, W_r)$. S is the confidence score of each location, measuring not only the degree of similarity between the object in the scene and the target object, but also the chance of a gripper pose to grasp the target object.

$S = (s_{r_0}, s_{r_1}, ..., s_{r_{n-1}}, s_{w_0}, s_{w_1}, ..., s_{w_{m-1}})$ is used to infer the location and gripper pose for the target object, where n is the number of defined gripper orientations, m is the number of defined gripper opening widths, and s is the confidence score of a grasp primitive. Details of using confidence score to infer the location of the target object are in Section *G*. To infer the gripper pose for the target object, we should compare the confidence score of each grasp primitives and find which one has the largest value. In other words, the gripper orientation for the target object is $\frac{180}{n} \times argmax(s_{r_0}, s_{r_1}, ..., s_{r_{n-1}})$ degrees, and the gripper opening width is $\frac{W_{max}}{m} \times (argmax(s_{w_0}, s_{w_1}, ..., s_{w_{m-1}}) + 1)$, where the W_{max} is the max opening width of the gripper. In our work, we set $n = 6, m = 3$, and $W_{max} = 96$ pixels.

B. Background: Model-Agnostic Meta-Learning

The model-agnostic meta-learning (MAML) [7] aims to learn the weight θ of a model f_θ, such that with one or more gradient descent using a small number of training data, the model can have a good performance on the new task, without overfitting. When adapting to a new task τ_i sampled from $p(\tau)$, the model's parameters θ are updated by a gradient descent with inner learning rate η as follows:

$$\phi = \theta - \eta \nabla_\theta L(f_\theta, \tau_i). \tag{2}$$

In the outer loop of the training phrase, the adapted-policy f_ϕ is evaluated on query examples from task τ_i. As shown in Fig. 3, the performance on the query examples is used to optimize the initial policy f_θ with meta learning rate β as follows:

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum L(f_\phi, \tau_i). \tag{3}$$

C. Basic Loss Function

We add a modulating factor proposed by [27] to binary cross-entropy loss to prevent our policy from being overwhelmed by easy samples when updated (e.g., background and the contour edge of a big object are easy samples; graspable portions of distractor objects and target objects are hard samples). To avoid confusion between o_1 and o_2, we use a weighting factor ν in the meta loss to make the network pay more attention to distinguishing o_1 and o_2. Finally, for each

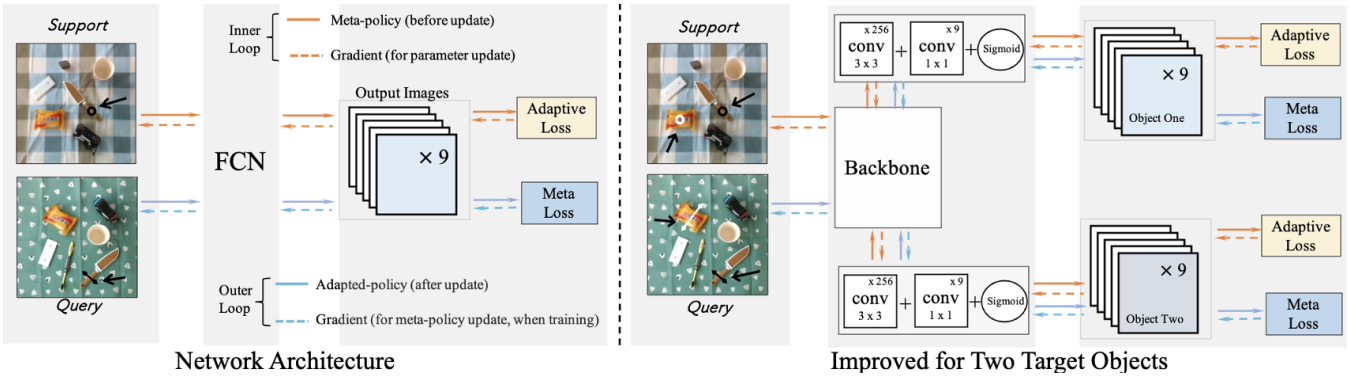

Fig. 3. The architecture of IGML. Left is our network architecture, and we augmented it to predict two instances by adding an identical branch (right). In the inner update (fast adaptation), the meta-policy is fine-tuning according to the adaptive loss. In the outer loop, the adapted policy is evaluated by the meta loss. We use a circle in the supporting image to represent the target grasping point and a line in the query image to represent a 4-DoF instance grasping.

pixel in the output images, the basic loss function is defined as:

$$L(\hat{y}) = \begin{cases} -\kappa \hat{y}^2 \log(1-\hat{y}), & if \text{ y=0} \\ -\alpha(1-\hat{y})^2 \log(\hat{y}), & if \text{ y=1} \\ -\nu \hat{y}^2 \log(1-\hat{y}), & if \text{ y=-1} \end{cases} \tag{4}$$

where κ, α and ν are weighting factors, y and $\hat{y}$ are ground truth and predicted value respectively. In the label of o_1 in query examples, o_1 with correct end-effector pose is labeled to 1, o_2 is labeled to -1, and others are labeled to 0. The same is done to generate labels for o_2. The labels of the support examples can be the same as the query examples, which not only provide the location of the target object but also provide the grasping pose. Also, we can only provide the target object's location in the supporting examples. Although the adaptive loss and meta loss share the same basic loss function, they differ in weighting factors significantly. In our work, we let $(\kappa = 1, \alpha = \nu = 9)$ in the meta loss, and $\nu = 0$ in the adaptive loss. Before the following discussions, it's worth mentioning that the total loss function is the average of the losses of all output pixels:

$$L = \sum_{i=1}^{n} L(\hat{y}_i)/n. \tag{5}$$

Therefore, the weighting factors κ and α in Eq. (4) determine the proportion of the positive sample in the total adaptive loss, thus significantly affecting the gradient descent in the inner loop.

D. Adaptive Loss and Meta Loss

A training iteration consists of two phases: meta-policy learns from supporting examples, and adapted-policy is evaluated by query examples, corresponding to the inner and outer loops, as shown in Fig. 3. For robotic instance grasping, the inner objective is to recognize the target object or recognize the target object and distinguish it from d_{sup}. The outer objective is to identify the target object, distinguish it from distractor objects, and master the ability to predict gripper pose. Intuitively, the inner objective is different from the outer objective, therefore we propose to use an adaptive loss in the inner loop to facilitate the inner objective and a meta loss in the outer loop to evaluate the outer objective. Besides, the meta-policy should master the ability to predict gripper poses. After the policy learns to recognize the target object during fast adaptation, it is capable of handling both recognition and grasping of the target object. In the following, we discuss the adaptive loss's weighting factor κ and α, the weighting factors for negative and positive samples, respectively.

E. Learning from Positive or Negative

Learning from only positive. To efficiently learn from the supporting examples, meta-policy can pay attention to the target object and the target grasping point during fast adaptation. Considering the Eq. (4), we let the weighting factor $\kappa = 0$ in the adaptive loss, so the gradient descent would only be caused by the positive samples. In this way, the meta-learning objective can be understood as: considering only the positive samples, such that after fast adaptation, the adapted policy can distinguish whether an object is the target object and can grasp the target object at the specified grasping point.

Learning from both positive and negative. The meta-policy can also learn from both positive and negative samples with distracted attention during fast adaptation. Intuitively, learning from negative examples would help distinguish the target object from d_{sup}. Considering the Eq. (4), both positive samples and negative samples should contribute to the total loss L, such that the adapted policy would boost the feature of the target object and suppress the feature of d_{sup}. In this way, the meta-learning objective can be understood as: considering both positive samples and negative samples in the supporting examples, such that after fast adaptation, the adapted policy can distinguish the target object from distractor objects, especially distinguish d_{sup}.

F. Model Architecture and Inference

Our network architecture takes the first 40 layers of ResNet-50 as the backbone, followed by two identical branches to output images for o_1 and o_2. Besides, we only need to use one branch to train a meta-policy that learns instance grasping for one target object. Our model takes in an RGB or RGB-D image of size 448×448 and outputs eighteen images of

TABLE I
THE SUCCESS RATE OF INSTANCE GRASPING. THIS TABLE USES N, C, AND D TO REPRESENT NON-CLUTTERED, CLUTTERED, AND DENSELY CLUTTERED ENVIRONMENTS.

Methods	Shots	Test Set 1			Test Set 2			Noise	Trained
		N	C	D	N	C	D		
IGML-PN	1-shot	78.3%(47/60)	63.3%(38/60)	8.3%(5/60)	93.3%(56/60)	91.7%(55/60)	6.7%(4/60)	50.0%	97.7%
IGML-PN	5-shot	81.7%(49/60)	66.7%(40/60)	8.3%(5/60)	**95.0%**(57/60)	**91.7%**(55/60)	6.7%(4/60)	61.1%	–
IGML-P	1-shot	88.3%(53/60)	83.3%(50/60)	75%(45/60)	86.7%(52/60)	83.3%(50/60)	70%(42/60)	77.8%	97.2%
IGML-P	5-shot	**95.0%**(57/60)	**86.7%**(52/60)	**76.7%**(46/60)	91.7%(55/60)	90.0%(54/60)	**80.0%**(48/60)	**88.9%**	–

TABLE II
THE SUCCESS RATE OF INSTANCE GRASPING IN NON-CLUTTERED SCENES.

Methods	Simulation(N)	Real(N)
CCAN [3]	83.9%	82.7%
ABRG [4]	87.6%	80.3%

size 28×28, from which we can infer the location and grasp pose for o_1 and o_2. We use models with RGB-D input in the following experiments since it is more robust to light and background.

The instance grasping inference process consists of three steps. First, we compare each pixel in the same position of images representing gripper orientation and take all the largest values to get the orientation confidence map. The same is done to get an opening width confidence map. Second, we take the average of these two maps to get a confidence map. According to the position of the maximum value in the confidence map, we can infer the location of the target grasping point of the target object. Finally, we return the target location to output images representing gripper orientations and gain the gripper orientation by inferring which image has the largest score in this pixel location. The same is done to infer gripper opening widths.

IV. EXPERIMENTS

Our experiments aim to address four main questions: 1) Does IGML allow a robot to learn from a small number of examples to recognize previously unseen objects and grasp them at specified grasping points with the desired accuracy? 2) Is IGML valid in object-rich or cluttered scenes? 3) Can our method improve the performance in different scenarios by learning from positive or negative samples? 4) Is teaching robots to grasp at the specified grasping point effective compared with indiscriminate grasping?

A. Datasets

Training Set. We collected a dataset containing 90 instance grasping tasks with 122 distinct objects, most of which are household objects and toys. Each instance grasping task contains 41 examples to grasp the same two objects orderly in different scenarios. Besides, the grasping point of the target object is the same in each task. The images are taken by Realsense D435, and we labeled them by a computer interface

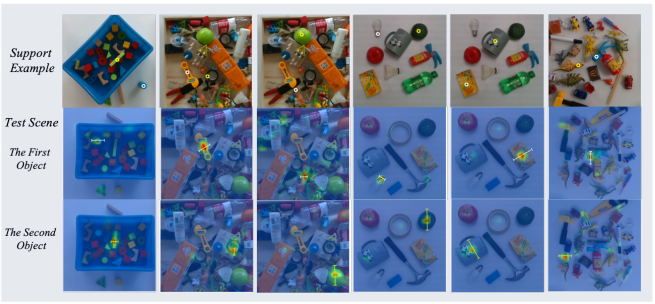

Fig. 4. Visualization of the results in cluttered scenes. The first row is the supporting examples. The second and the third rows are inferred confidence maps of the first and second objects in new environments.

that enables interaction with images by mouse. We did the translation, rotation, flip, brightness, and color augmentations to enrich the dataset and interchanged the order of the first and second target objects in an example when training. We leave two examples in each instance grasping task to test our model's performance.

Test Sets. We collected two test sets. Test Set 1 assumes that distractor objects in the test scene may not belong to d_{sup}. Test Set 2 has an assumption that distractor objects in the test scene all belong to d_{sup}. Each test set has 30 instance grasping tasks with a total of 180 images, which consists of (i) 10 tasks in non-cluttered environments (e.g., Fig. 2), (ii) 10 tasks in object-rich or cluttered environments (e.g., Fig. 4), and (iii) 10 tasks in dense clutter (e.g., Fig. 4). It's worth mentioning that this dataset restricts the densely cluttered scenes to constrain that the target objects are graspable, visible, and without much occlusion. We included 47 novel objects in these two test sets, most of which are household objects and toys. Besides, three tasks with environmental noise, such as shadows and depth noise, are collected to test our model's robustness.

Evaluation Metric. We consider the position and gripper opening width error does not exceed two grids (7.1% of the image), and the rotation error does not exceed 24° as a success of grasping the target instance at the specific grasping point. Note that we consider success only when two target objects are both correct and without confusion.

B. Results on Test Sets and Discussion

We trained a meta-policy that learns from only positive samples (IGML-P) and a meta-policy that learns from both

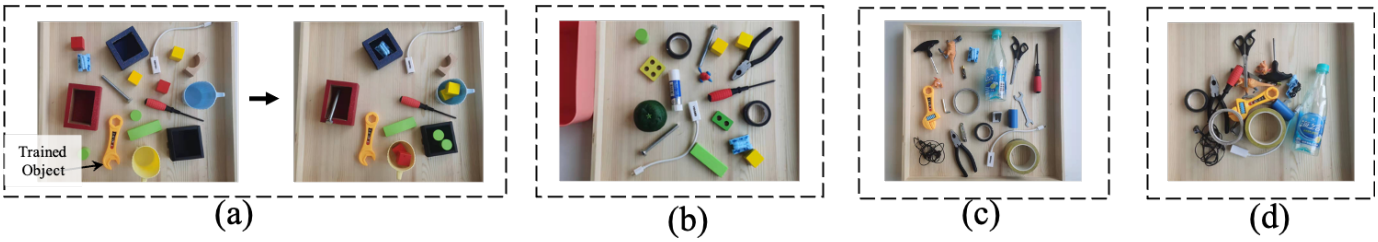

Fig. 5. Illustration of (a) pick-and-place scenario with seventeen novel objects and one trained object, (b) pick-and-place scenario with eighteen novel objects, (c) objects for indiscriminate grasping experiments, (d) heavy clutter scene for indiscriminate grasping.

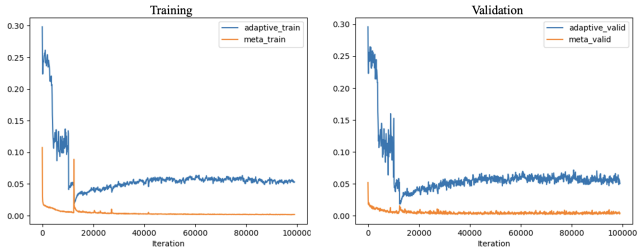

Fig. 6. The curves of the adaptive loss and meta loss.

positive and negative samples (IGML-PN). The curves of meta loss and adaptive loss of IGML-P are shown in Fig. 6. Table I shows the performance of IGMLs in non-cluttered environments, object-rich or cluttered environments, densely cluttered environments, scenes with environmental noise, and scenes with trained objects. In comparison, Table II shows the original success rates of state-of-the-art instance grasping methods CCAN [3] and ABRG [4]. CCAN was trained with simulated data, including 135 training objects, and tested with 15 novel objects. As CCAN and ABRG are trained with simulated data, we compare our test results with their test results in simulation environments, in which CCAN achieved a success rate of 83.9% and ABRG achieved a success rate of 87.6%. In comparison, IGML-P shows a success rate of 95.0% in Test Set 1 and 91.7% in Test Set 2, and IGML-PN shows a success rate of 95.0% in Test Set 2. Considering the evaluation metric of our work is more difficult than prior works, the results indicate that IGML-P can significantly surpass state-of-the-art methods, demonstrating our framework's effectiveness. The results also show that IGML-PN can significantly outperform state-of-the-art methods in scenarios where distractors all belong to d_{sup}, which demonstrates the effectiveness of learning from both positive and negative samples in special cases.

Comparing the results of 1-shot and 5-shot, we found that the success rate of instance grasping is greatly improved when multiple adaptation images are provided. We consider that a possible reason is that fast adaptation with a single adaptation image means that the meta-policy needs to learn to recognize the target object from a single visual angle, which will lead to an insufficient understanding of the target object. In contrast, multiple adaptation images enable the meta-policy to recognize the target object comprehensively.

IGML-P versus IGML-PN. Test Set 1 shows that the performance of IGML-P can significantly surpass IGML-PN in common cases, and IGML-P is more robust to environmental noise. In comparison, the performance of IGML-PN is poor in Test Set 1 since distractor objects that do not belong to d_{sup} but have colors or shapes similar to the target object tend to get high values in the confidence map. These results demonstrated the effectiveness of learning from only positive samples, which enables the meta-policy to learn efficiently from the supporting examples. In addition to common scenarios, we also considered a special case where distractors all belong to d_{sup}. The results of IGML-PN on Test Set 2 demonstrate the effectiveness of learning from both positive and negative samples in special cases.

C. Pick-and-place Scenarios

Pick-and-place is a vital application of robotic instance grasping. We consider the success rate of IGML-P is high enough to perform pick-and-place in cluttered scenarios. Therefore, we build two physical world pick-and-place scenarios to test the performance of IGML-P, as shown in Fig. 5. For statistical convenience, we consider a success if a target instance is grasped successfully rather than grasping at the specific portion of the target object.

Setup. Experiments are carried out in a new background with mostly novel objects. The goal of the first scenario is to pick up five kinds of target objects (e.g., o_1) with multiple instances in the scene out of the clutter and place them into their corresponding boxes (e.g., o_2, we take a box as an object). The goal of the second scenario is to pick-and-place objects in the scene in a predefined order. The UR5 robot arm handles these pick-and-place scenes by dynamically learning ($\sim$60ms, 5-shot) and performing new instance grasping tasks.

Results. We take the confidence score equal to 0.8 as a boundary to decide whether it is a target object. Then we tested ten times for each scenario, achieving a precision rate of 93.4% and a recall rate of 87.1%. The result demonstrates the practicality of IGML-P, which significantly surpasses state-of-the-art methods. Besides, this experiment also depicts that our model can generalize to predict the locations of novel boxes/containers.

D. Indiscriminate Grasping Experiments

Along with the state-of-the-art results in robotic instance grasping, we also demonstrate that our meta-policy (initial

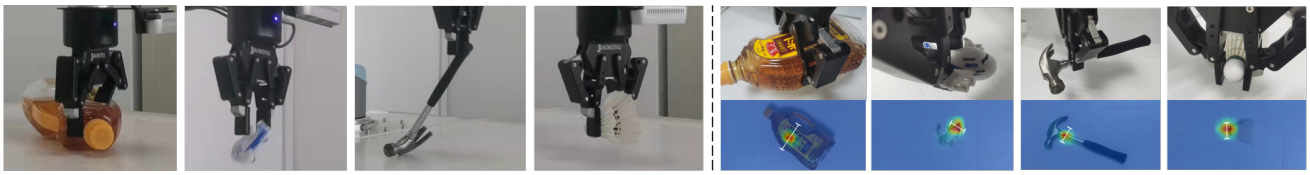

Fig. 7. Illustration of indiscriminate grasping compared with teaching robots to grasp at a specified grasping point. Left are typical failures caused by a state-of-the-art indiscriminate grasping model in the experiments, and they are corrected when we teach robots where to grasp (right).

model) can predict indiscriminate grasps for previously unseen objects. As shown in Fig. 5, we used 20 household objects and toys to evaluate the performance of our meta-policy in the physical world using the UR5 robot arm and a novel background. Each object was tested ten times with random location and orientation. The robot performed 185 successful grasps of the total 200 attempts, achieving a success rate of 92.5%. The result demonstrates that the meta-policy has mastered the ability to grasp, showing a competitive performance against state-of-the-art methods (e.g., [8], [10] and [11] achieved a success rate of 89%, 92%, and 95.4%, respectively). In addition, we also tested the performance of our meta-policy in heavy clutter ten times. As a result, the robot takes an average of 23.5 attempts to grasp the 20 objects, achieving a success rate of 85.1%.

E. Grasp objects with complex shapes

To verify the effectiveness of teaching robots to grasp at a specified grasping point, we carried out grasping experiments in more challenging scenarios containing novel objects with complex shapes and/or uneven mass distributions, as shown in Fig. 7. Each object was tested one by one twenty times with a random location and pose. For the seasoning bottle, the difficulty of grasping is that there are only few appropriate grasping points on the skeleton of the middle part of the bottle, and other grasping points may cause a collision or other failures. For the toy airplane, the difficulty lies in the complex geometry. But due to its small size, a poor grasp pose may also pick it up. The difficulty of hammers and badminton has been mentioned earlier. We compare our results with the state-of-the-art indiscriminate grasping method GR-ConvNet [11], which has a much higher performance in grasping than CCAN [3] and ABRG [4]. In the experiment of grasping badminton, GR-ConvNet often failed to grasp its head (e.g., we call it a failure) and also failed to pick it up, as GR-ConvNet often located the tail of the badminton and generated a poor gripper orientation. Table III shows the success rate of IGML-P compared with GR-ConvNet. The results demonstrate that teaching the robot to grasp at a specified grasping point leads to reliable grasping results, even for objects with complex shapes.

TABLE III
THE SUCCESS RATE OF GRASPING OBJECTS WITH COMPLEX SHAPES

Methods	Hammer	Badminton	Airplane	Bottle
GR-ConvNet [11]	10/20	6/20	16/20	1/20
IGML-P	20/20	20/20	20/20	16/20

F. Other Manipulations

Along with the state-of-the-art results in test sets and the physical world, we also show that our IGML has the potential to perform a series of manipulations. We take two representative manipulation tasks as examples. The first is the bottle placement task from [26], which aims to grasp novel bottles and place them upright. The second is the insertion task from [22], which uses trained objects for experiments.

Bottle placement task. To place bottles upright, robots need to distinguish the head and tail of bottles or can identify the pose of a bottle. Therefore, we propose a two-point method to extend pose-independent grasping, which predicts gripper orientation in the range of [0, 180) degrees [10], [11], to pose-relevant grasping, which predicts object orientation in the range of [0, 360) degrees. As shown in Fig. 8 (a), we use the two heads of IGML to learn two grasping points of an object, namely p_1 and p_2, so we can get a vector v_p related to the object's pose. We use p_1 as the grasping point, v_p as the orientation vector of the gripper so that the robot can infer the relative pose of the grasped object to the gripper. We let p_2 be the head of the bottle or mug and set a placement pose for the robot to achieve the bottle placement task, as shown in Fig. 8 (b) and Fig. 8 (c).

We performed twenty tests on three bottles and one mug in scenes with two or more distractor objects. As a result, we achieved a success rate of 95%, showing that our robot can perform non-trivial placements for novel objects. Although we only performed the bottle placement task, the two-point method can actually generalize to more tasks, such as flipping a cup in a two-step approach like [26], inserting a screw into a hole, or tool manipulation [19]. Compared with previous work about non-trivial placements, the advantage of our method is that IGML can recognize novel objects and is not limited to a certain object class.

Insertion task. To perform the insertion task (shown in Fig. 8 (d)), we finetuned IGML with 20 gradient steps during fast adaptation to improve the localization accuracy of grasping and inserting points. Following [22], we conducted twenty single-hole insertions with three distractors placed in the test scene and achieved a success rate of 80%. The results show that we have comparable performance against [22], which demonstrates the effectiveness of teaching robots to grasp at a specified grasping point.

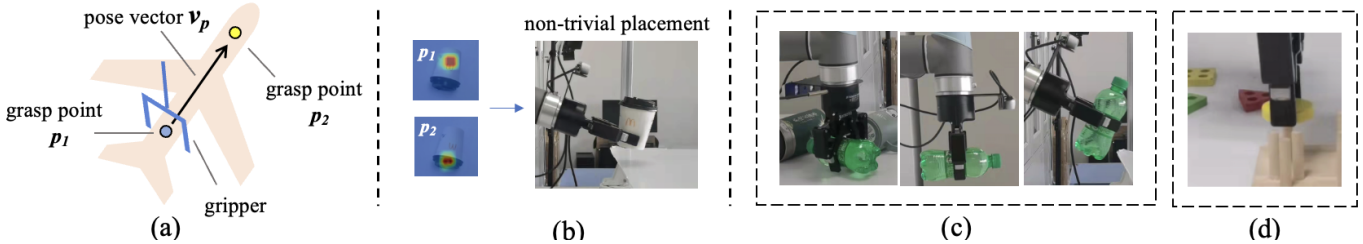

Fig. 8. (a) an example of two-point pose-relevant grasping. (b) an example of applying the two-point method to place a cup upright. (c) the process to place the bottle upright. (d) an example of an insertion task.

V. Conclusion

We presented IGML, a novel and practical instance grasping meta-learning framework, which can fast adapt to grasp the specific portion of the novel target object out of the clutter. We show the study of learning from positive or negative samples in the supporting examples to achieve a higher success rate in different scenarios. Experimental results on the test sets and the physical world demonstrate the effectiveness and practicality of IGML, which significantly surpasses state-of-the-art methods.

Limitations. First, we implicitly assumed that the visual angle of the camera for the supporting examples and test scene should not be distinct too much (e.g., a supporting example collected on a top-down visual angle wouldn't enable the robot to recognize and grasp the target object robustly on a front-back visual angle). Second, the difficulties posed by complex new environments haven't been considered in this work.

References

[1] K. Fang, Y. Bai, S. Hinterstoisser, S. Savarese, and M. Kalakrishnan, "Multi-task domain adaptation for deep learning of instance grasping from simulation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3516–3523.

[2] D. Park, Y. Seo, D. Shin, J. Choi, and S. Y. Chun, "A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7300–7306.

[3] J. Cai, X. Tao, H. Cheng, and Z. Zhang, "CCAN: Constraint co-attention network for instance grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8353–8359.

[4] Y. Yang, Y. Liu, H. Liang, X. Lou, and C. Choi, "Attribute-based robotic grasping with one-grasp adaptation," *arXiv preprint arXiv:2104.02271*, 2021.

[5] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[6] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020.

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

[8] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[9] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[10] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International journal of robotics research*, vol. 39, no. 2-3, pp. 183–201, 2020.

[11] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9626–9633.

[12] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3750–3757.

[13] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 7223–7230.

[14] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine, "End-to-end learning of semantic grasping," *arXiv preprint arXiv:1707.01932*, 2017.

[15] E. Jang, C. Devin, V. Vanhoucke, and S. Levine, "Grasp2vec: Learning object representations from self-supervised grasping," *arXiv preprint arXiv:1811.06964*, 2018.

[16] J. Cai, H. Cheng, Z. Zhang, and J. Su, "Metagrasp: Data efficient grasping by affordance interpreter network," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4960–4966.

[17] Y. Fleytoux, A. Ma, S. Ivaldi, and J.-B. Mouret, "Data-efficient learning of object-centric grasp preferences," 2021.

[18] F. Hélénon, L. Bimont, E. Nyiri, S. Thiery, and O. Gibaru, "Learning prohibited and authorised grasping locations from a few demonstrations," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 1094–1100.

[19] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

[20] E. Matsumoto, M. Saito, A. Kume, and J. Tan, "End-to-end learning of object grasp poses in the amazon robotics challenge," *Advances on Robotic Item Picking*, pp. 63–72, 2020.

[21] R. Jonschkowski, C. Eppner, S. Höfer, R. Martín-Martín, and O. Brock, "Probabilistic multi-class segmentation for the amazon picking challenge," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1–7.

[22] L. Berscheid, P. Meißner, and T. Kröger, "Self-supervised learning for precise pick-and-place without object model," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4828–4835, 2020.

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

[24] A. Bonardi, S. James, and A. J. Davison, "Learning one-shot imitation from humans without humans," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3533–3539, 2020.

[25] S. James, M. Bloesch, and A. J. Davison, "Task-embedded control networks for few-shot imitation learning," in *Conference on Robot Learning*. PMLR, 2018, pp. 783–795.

[26] M. Gualtieri, A. ten Pas, and R. Platt, "Pick and place without geometric object models," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 7433–7440.

[27] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.