Discriminative Active Learning for Robotic Grasping in Cluttered Scene

Boyan Wei*, Xianfeng Ye*, Chengjiang Long, Zhenjun Du, Bangyu Li, Baocai Yin, and Xin Yang

***Abstract*— Robotic grasping is a challenging task due to the diversity of object shapes. A sufficiently labeled dataset is essential for the grasp pose detection methods based on deep learning. However, data annotation is a costly procedure. Active learning aims to mitigate the greedy need for massive labeled data. In this work, we propose a Discriminative Active Learning (DAL) framework for robotic grasping algorithms. DAL is an effective strategy that utilizes a shared encoder to derive latent features from both labeled data and unlabeled data. A discriminator is established to estimate the informativeness of each unlabeled data sample and decide whether they should be annotated for the next epoch. Moreover, an annotation interface is also developed to annotate the chosen data. We evaluate DAL with real-world grasp datasets and show superior performance, especially when the amount of labeled data is little. Considering annotation noise, we perform an experiment on a noisy dataset and demonstrate that our proposed framework is stable to noisy annotation. Besides, we train a model with about 60% data selected by DAL of the whole dataset and it can still handle a real-world grasp detection task in cluttered scene on a real robot.**

***Index Terms*— Robotic grasping, Active Learning, grasp pose detection**

I. INTRODUCTION

The technique of adaptive robotic grasping is an essential function for a service robot in the real environment. When a robot is placed in a new room surrounding by a set of novel objects, as shown in Fig. 1, it either relies on existing learned or designed knowledge to detect grasps, or learns how to execute a successful grasp from new knowledge. Prior knowledge is not always reliable in real-world cases, thus learning is the necessary option to adapt a robot to a new environment and achieve a higher grasp success rate. However, learning is an extremely resource-consuming process in an unexploited environment. As for deep learning, data labeling consumes the most resources including human resources and time. As is shown in Table. I, taking the Jacquard grasp dataset as an example, the budget of annotation is extremely large. It is not realistic to annotate thousands of data for a specific grasp task, thus we are looking for a way to train a satisfactory grasp detection model with as few labeled data as possible.

Fig. 1. Kinova MOVO with some novel objects in cluttered scene. Using prior knowledge, it may fail to generate good grasp poses.

Research on antipodal robotic grasping has developed significantly in the last two decades [1], [2], [3], [4], [5]. Thanks to the development of deep learning, recent researchers [6], [7], [8], [9], [10], [11], [12], [13] have made huge advancements in grasp pose detection. Some researchers [6], [7] adapt object detection algorithms to grasp detection, while there are some regression-based approaches [14], [8], [15], [11], [12], [13] predict grasps directly without the region proposal stage. However, deep learning is always hungry for data. Without sufficient data, deep learning based approaches are hard to learn accurate knowledge and provide satisfactory results.

TABLE I

ANNOTATION BUDGET OF JACQUARD GRASP DATASET

Objects amount in Jacquard	**Ground truth value for each annotation**	**Annotation times for each picture**
11000+	4	10
Annotation times for all	**Time for each ground truth value**	**Time of labeling the whole dataset**
2200000	1 second	611 hours/labeler

Annotation for datasets is an exhausting and costly but necessary procedure. Data collection and annotation are essential for better performance when we adapt a trained model to a new environment. Inspired by active learning [16], researchers have proposed a series of effective methods [17], [18], [19], [20], [21], [22] to reduce the cost of data labeling by selectively choosing data to label rather than labeling the whole dataset. The core idea of active learning is that the most informative data sample would contribute more to improve model performance than other random samples. Generally, current active learning strategies can be categorized into diversity-based approaches and uncertainty-

This work was supported in part by National Key Research and Development Program of China (2022ZD0210500), the National Natural Science Foundation of China under Grant 61972067/U21A20491/U1908214, and the Innovation Technology Funding of Dalian (2020JJ26GX036).

[1]Xianfeng Ye, Boyan Wei, Baocai Yin, Xin Yang are with the School of Computer Science and Technology at Dalian University of Technology, Dalian, 116024, China. (xiangjiaopi@mail.dlut.edu.cn; swaggerwey@mail.dlut.edu.cn; ybc@dlut.edu.cn; xinyang@dlut.edu.cn)

[2]Chengjiang Long is with JD Finance America Corporation, 675 E Middlefield Rd, Mountain View CA, USA. (cjfykx@gmail.com)

[3]Zhenjun Du, Bangyu Li are with Shenyang SIASUN Robot and Automation Co. Ltd, ShenYang, 110168, China. (duzhenjun@siasun.com; libangyu@126.com)

* Boyan Wei and Xianfeng Ye are the joint first authors. Xin Yang and Chengjiang Long are the joint corresponding authors.

based approaches. Diversity-based approach [23], [24], [22] selects the most representative data samples of the unlabeled pool, while uncertainty-basedd approach [25], [18], [19], [20] focuses on how to define a criterion to measure the uncertainty of data samples.

Active learning has shown its efficiency in several tasks such as semantic segmentation [24], [18] and object detection [25], [26]. However, there are two restrictions when an active learning strategy is applied on robotic grasping tasks: (1) Grasp pose detection methods are based fully or partially on regression networks. Therefore, active learning strategies designed for classification tasks are not suitable, since posterior probabilities from classification networks are necessary for these strategies. (2) The number of parameters in grasp pose detection methods is relatively small in consideration of real-time performance, which means an active learning strategy with a huge number of parameters would result in unnecessary resource consumption. In consideration of the restrictions, it is difficult to adapt current active learning strategies to robotic grasping tasks.

In this paper, we propose an active learning framework for grasp pose detection algorithms. The main contributions of this paper are summarized as:

- We propose a discriminative active learning framework that is suitable for grasp pose detection algorithms. Besides, we develop an interface in our proposed framework to efficiently and conveniently annotate selected data samples.
- The proposed active learning framework shares the encoder with a grasp pose detection network and takes full advantage of both labeled data and unlabeled data. A discriminator is established to utilize the latent features extracted with the shared encoder and estimate the distance from each unlabeled data sample to the labeled pool. With the results of estimation, we can selectively choose more informative data samples to train the grasp detection network.
- We evaluate the proposed method on real-world grasp datasets, the Cornell Grasp Dataset and the Jacquard Grasp Dataset, and demonstrate that our method achieves superior performance, especially when the size of the labeled pool is relatively small. Moreover, we establish a noisy Cornell Dataset. to simulate annotation noise in real life. Our proposed active learning framework still shows stable performance even though annotation noise exists.
- We demonstrate that the model trained with our proposed active learning framework can be deployed on a real robot to perform real-time grasp pose detection.

II. RELATED WORK

Robotic Grasping. Grasp pose detection is a task to generate a stable robotic grasp pose for a given object. It has been researched for decades resulting in a wide range of approaches [1], [9], [10], [27], [6], [7], [8], [12], [28]. Current robotic grasping methods can be categorized into two types: analytic methods and empirical methods. Analytic methods [1], [27] calculate grasps with physical models, kinematics and dynamics, while empirical methods [6], [7], [8], [12], [9], [10], [11], [13], [28] take advantage of human labels and experience-based approaches.

Empirical methods are attentive to learning from experience. With the enormous development of deep learning, a great number of neural networks [6], [7], [8], [9], [10], [28], [29], [30] were designed to deal with grasp pose detection. Most of these techniques follow a common pipeline: generate grasp candidates from inputs (images or point clouds), score each grasp candidate and output the grasp configuration with the highest score. Several approaches [6], [7], [28] used a sliding window as object detection algorithms to procure grasp candidates which tends to cause unnecessary resource consumption. Approaches [29], [30] transform depth data into point cloud and feed it into PointNet++ mostly are costly on computational time. To reduce execution time, some approaches [6], [31] pre-processed and pruned grasp candidates to be less time-consuming, but result in discarding some potential grasps.

Instead of adaptation of other object detection algorithms, some researchers proposed regression methods [8], [15], [14], [32], [12] for grasp pose detection. Kumra *et al.* [14] and Redmon *et al.* [32] used a deep convolutional neural network to regress a single grasp for each input image. Morrison *et al.* [8] proposed a fully convolutional network that can generate grasp configurations directly without sampling grasp candidates. Kumra *et al.* [15] utilized the residual convolutional neural network to extract more features without much loss of real-time performance.

In spite of the excellent performance current deep-learning methods can achieve, they are still anxious for sufficient labeled data to train a satisfactory model. Furthermore, when a model trained on existing datasets performs badly on a set of new objects, a new dataset needs to be created, which is an expensive procedure on both time and human resources.

Active Learning. Active learning [33], [34], [35], [36], [37] aims to reduce the cost of data annotation by selectively choosing the most representative data rather than the whole dataset to be labeled. Broadly, there are two kinds of active learning strategies: diversity-based strategy and uncertainty-based strategy.

Diversity-based active learning strategy [38], [39], [40], [23], [24], [22] concentrates on the distribution of data and aims to choose the most informative data. A naive way to deduce diversity is running a clustering algorithm on unlabeled data [40]. Ebert *et al.* [23] proposed Graph Density which is calculated with Manhattan distance to construct a graph structure and updated when some instances are removed from the unlabeled pool. The instance with the highest graph density will be added to the labeled pool and takes part in the next training iteration. Obviously, diversity-based strategy is task-agnostic. Because it assumes the instances with the highest graph density as the most informative data. However these instances may not actually benefit the improvement of the model.

Uncertainty-based active learning strategy [41], [42], [43],

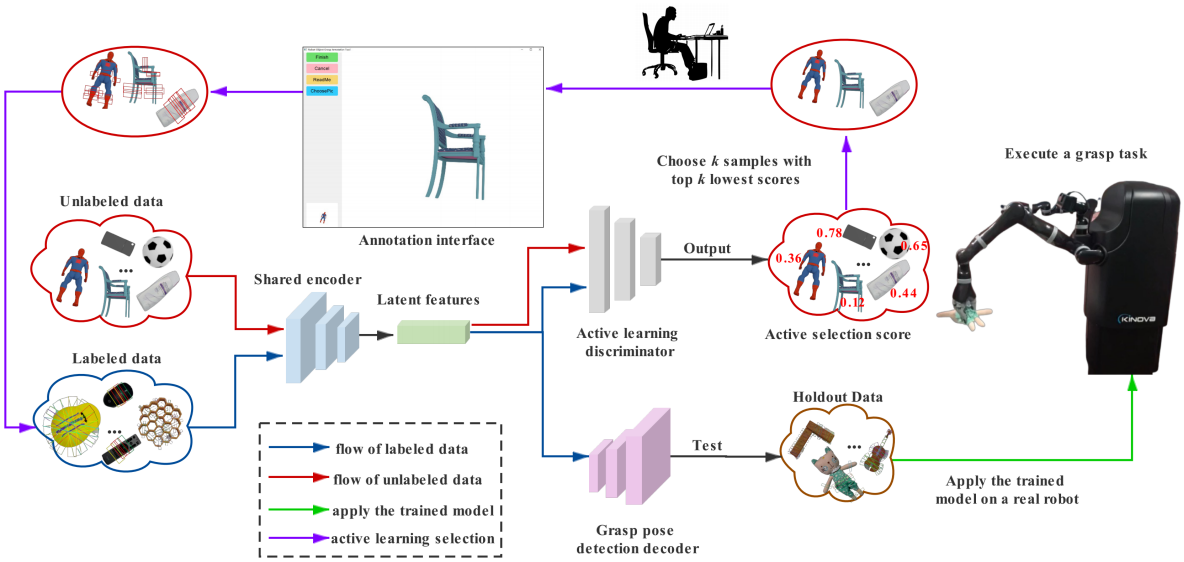


Fig. 2. The whole structure of our proposed active learning framework DAL for robotic grasping tasks. DAL takes the following steps to select data and train the grasp detection network: (1) The active learning discriminator and the grasp pose detection decoder utilize a shared encoder that takes both labeled data and unlabeled data as input. (2) The active learning discriminator estimates a score for each unlabeled data with the shared latent features. (3) The k samples with the lowest top k scores will be chosen and input into our designed annotation interface. (4) After annotation, the selected data will be removed from the unlabeled pool and added into the labeled pool. (5) The grasp pose detection decoder is trained with the latent features of the labeled data. (6) Finally, the grasp detection model including the shared encoder and the grasp pose detection decoder will be applied on a real robot when it achieves a satisfactory performance on the holdout data.

[44] looks for the data that a model is most uncertain about. Uncertainty-based strategy usually defines a metric of uncertainty that can be derived from deep-learning models. Researchers have found plenty of ways to define uncertainty so far. Uncertainty Sampling [33] utilized the posterior probability to measure uncertainty. Another approach called QBC (Query-By-Committee) [42] trained an ensemble of models simultaneously and the disagreement among these models is defined as uncertainty. Nevertheless, an insufficiently trained model is likely to bring about the inaccurate estimation of uncertainty.

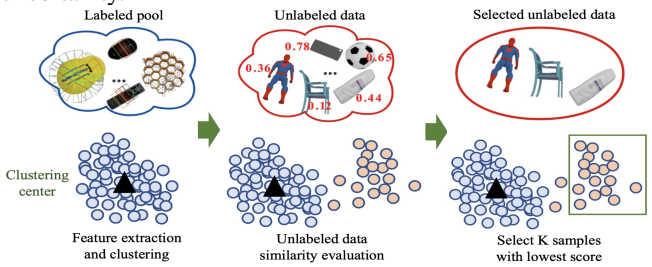


Fig. 3. Active learning theory explanation: take clustering active learning strategy as an example, the discriminator can tell the distance between unlabeled data and labeled pool, and select the data with the lowest score to annotate them.

Active learning strategies designed for classification tasks have an extremely limited application on regression tasks due to the indispensability of posterior probabilities. As illustrated in Fig. 3, active learning method can discriminate the difference among data. If the model has already acquired a sample, and then another similar sample comes up, active learning method can detect the similarity between them, and decide not to annotate the redundant unlabeled sample. Recently, some researchers focus on task-agnostic uncertainty-based active learning strategies. Yoo *et al.* [20] constructed a loss prediction module to predict loss of training data. A larger loss indicates that the model is more uncertain, hence the predicted loss represents uncertainty. However, it does not perform well in the early stage of active learning, since few labeled data results in a deficiently trained loss prediction module. Sinha *et al.* [19] used a reconstruction network to obtain features of input images and find the most unlabeled-like data sample. Nevertheless, it takes no advantage of the latent features extracted by backbone algorithms and a huge amount of extra parameters are necessary due to the reconstruction network.

Algorithm 1 Active learning for robotic grasping

Input: Labeled dataset D_L, unlabeled dataset D_U, test dataset D_T, expect accuracy e, maximum iteration K, accuracy optimized iteratively acc

Output: Trained Θ_{GSP} and Θ_{DAL}

$k \leftarrow 0$;
while $acc < e$ and $k < K$ **do**
  $k \leftarrow k + 1$;
  $\Theta_{GSP}, \Theta_{DAL} \leftarrow Train(D_L, D_U)$;
  $p(D_U) \leftarrow \Theta_{DAL}(D_U)$;
  $X_s^\star \leftarrow \text{argmin}_{X_s \in D_U} p(X_s)$;
  We ask human labelers to annotate $X_s^\star$ and get the annotation $Y_s^\star$;
  $D_L \leftarrow D_L \cup (X_s^\star, Y_s^\star)$;
  $D_U \leftarrow D_U - X_s^\star$;
  $acc \leftarrow Test(\Theta_{GSP}, D_T)$;
end while
return Θ_{GSP} and Θ_{DAL}.

III. APPROACH

As illustrated in Fig. 2, we propose a discriminative active learning framework for real-world robotic grasping problem. We collect a large data set including both labeled data D_L and unlabeled data D_U for training and a small hold-out data set D_T for evaluating. Initially, we randomly sample k data

samples to form D_L and ask human labelers to annotate. With the initial D_L and D_U, we can train a grasp model Θ_{GSP} and a discriminative active learning model Θ_{DAL}. With the output of the Active Learning Discriminator, we assume samples with the lowest top k scores as the most informative samples and ask human labelers to annotate them, and put the newly annotated examples into D_L. Then we can use the updated D_L and D_U to retrain the models Θ_{GSP} and Θ_{DAL} and select the most informative samples to annotate again. With this loop continues, we collect more and more labeled data in D_L and the size of D_U reduces gradually. Meanwhile, we can apply the learned Θ_{GSP} at each round to evaluate the performance with the real-word robots. Algorithm 1 shows how the active learning strategy works on robotic grasping tasks.

In the following, we are going to describe the details of our architectures for Θ_{GSP} and Θ_{DAL}, active selection strategy, the designed user-interface for annotation, the verification with our real-world robot and the implementation detail.

A. Joint Architecture for Grasp Detection and Discriminative Active Learning

As shown in Fig. 2, our framework mainly consists of two components, *i.e.*, the grasp model and the discriminative active learning model. To specify, the grasp model is trained with the labeled data to predict the grasping center, angle and width, while the discriminative active learning model is trained to produce the probability to distinguish between labeled and unlabeled data. The feature extraction is shared between these two models and both models are trained jointly so that we can make full use of both the labeled data and unlabeled data to ensure a solid feature extraction part.

In principle, any grasp detection network can be used as the grasp model in our framework. In this paper, we choose GG-CNN [8] as our backbone grasp pose detection network for evaluation for two reasons: (1) GG-CNN is a real-time grasp pose detection network with relatively high accuracy. (2) GG-CNN has a simple architecture and a small number of parameters, which allows us to concentrate on the implementation of our active learning framework. We shall emphasize that the GG-CNN here also employs the unlabeled data in the shared feature extract part, which is different from the origin GG-CNN.

For the active learning model, we use an MLP (multi-layer perceptron) as the discriminator to process the latent features and produce a probability to check whether the input example is likely to be a labeled example. The probability represents the similarity of a sample to be the labeled data. A larger probability denotes that the input sample is closer to labeled data, which indicates a lower priority of being annotated. On the contrary, samples with smaller probabilities should be chosen because it is more likely that the shared encoder has not been trained with these latent features yet. In this way, diverse samples are selected from the unlabeled pool and annotated, which avoids picking similar samples and helps the grasp detection network to obtain better training. Moreover, both labeled data and unlabeled data are utilized, thus abundant data samples are provided to train the shared encoder, which helps to improve the performance of the grasp detection network as well.

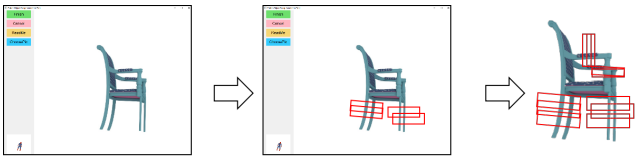


Fig. 4. The appearance and workflow of our designed interface. There are four buttons including "Finish", "Cancel", "Readme" and "ChoosePic" at the top of the left column. A thumbnail of the next data waiting to be annotated is placed at the lower left corner.

Our loss function for jointly modeling is formulated as:

$$\mathcal{L} = \mathcal{L}_{GSP} + \alpha * \mathcal{L}_{DAL} \tag{1}$$

where $\mathcal{L}_{GSP}$ follows the original loss functions in the grasp detection network GG-CNN, and $\mathcal{L}_{DAL}$ is defined as binary cross entropy (BCE) loss. We analyzed the performance of various combinations of the two loss functions and found that a factor α between 0 and 0.5 is proper is necessary to balance the different scales between them.

B. Active learning selection strategy

At each active learning iteration, we can train both the grasp detection model Θ_{GSP} and the discriminative active learning model Θ_{DAL}. We apply the trained model Θ_{DAL} on the unlabeled data and get the corresponding probabilities $p(D_U)$ measuring the likelihood of being labeled data. To select the most informative samples, intuitively we should choose the samples that are least likely to be the labeled data. Therefore, we can actively choose the samples by the following strategy:

$$X_s^\star = \text{argmin}_{X_s \in D_U} p(X_s). \tag{2}$$

Obviously, the sample with the least score will be chosen in a priority order since the discriminator estimates that its latent features are the most unfamiliar. We can choose $|X_s^\star|$ samples with top least scores according to the need. After a human labeler annotate the chosen data, a new round of training will be triggered.

C. Annotation interface

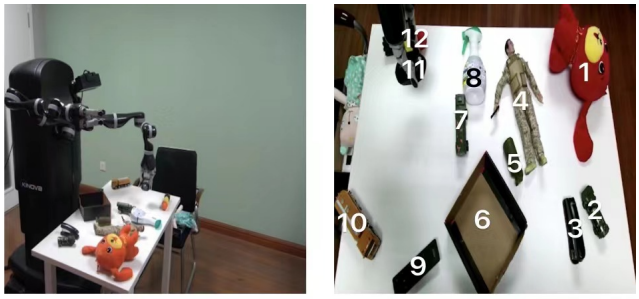
Annotation is an important procedure of our proposed active learning framework after data selection. A specific rectangle representation [45] consisting of the center point, the angle and the width of grasp pose is used in our framework. However, there are few available convenient grasp annotation tools. Therefore, we develop a user interface for grasp pose annotation and utilize it in our proposed active learning framework. As illustrated in Fig. 2, the data sample with the lowest active learning score is input into the interface and a human labeler can annotate the input data with the interface. After annotation, the selected data moves with its label from the unlabeled pool to the labeled pool and takes part in the next iteration of training.

Fig. 4 shows the workflow of the designed interface. The selected data is displayed in the middle and a human labeler can perform annotation on the image. A thumbnail is shown at the bottom of the left column to indicate whether there

are more selected data waiting for annotation. When all annotations are done, the finish button will trigger another round of training for the grasp detection network and the active learning strategy.

D. Application on a real robot

After several iterations of active learning selection, the trained model can achieve satisfactory performance on the holdout data as shown in Fig. 2. Then the trained grasp detection model is ready to be applied on a real robot and execute a grasp detection task. As shown in Fig. 5, the objects grasped are different in texture, shapes, materials, and scales.



(a) Kinova MOVO grasping an object.

(b) First-person perspective of Kinova MOVO.

Fig. 5. Real-world grasp task on Kinova MOVO using the grasp pose detection model trained by our proposed framework DAL.

E. Implementation detail

The experiments are conducted on the Ubuntu16.04 with Intel Xeon CPU E5-2650 and NVIDIA GeForce TITAN V GPU. All the algorithms are implemented in Pytorch [46]. In the training phase, the factor α in Equation 1 is set to 0.1. We use the Adam optimizer for the backbone GG-CNN as the original paper does, and the SGD optimizer for the active learning module. No data augmentation is performed in all the experiments.

IV. EXPERIMENTS

A. Datasets and metrics

We choose two grasp datasets, the Cornell Grasp Dataset [45] and the Jacquard Grasp Dataset [47], for training and evaluating our active learning framework. Both datasets are based on real-world objects.

The Cornell Grasp Dataset consists of 885 RGB-D images with a resolution of 640×480 pixels. There are 240 different real-world objects with 5110 positive and 2090 negative grasps in the dataset. We randomly select 100 images to form the initial labeled pool. For the backbone grasp pose detection network we choose, only positive grasps are necessary. Grasps are represented in the corner points' coordinates of grasp rectangles. It is a realistic dataset to evaluate our active learning framework because obtaining thousands of data is expensive in a real-world task.

The Jacquard Grasp Dataset is much larger than the Cornell Grasp Dataset, and composed of over 11k objects with about 5 RGB-D images for each object. It is a sufficient dataset for research on grasp pose detection, yet it would not be realistic to collect such a huge amount of data for a real grasp task. Therefore we randomly select 300 objects with 5 RGB-D images from the Jacquard Grasp Dataset and obtain a sub-dataset containing 1500 RGB-D images as the initial labeled pool. We reckon that if real-life conditions are taken into consideration, the sub-dataset with an appropriate amount of data is proper to evaluate the efficiency of our proposed active learning framework.

Aimed to show the performance of our active learning strategy, we selected five representative methods for comparison including RS (Random Sampling), GD (Graph Density [23]), LL (Learning Loss [20]), VAAL [19], CS(Core-set [22]) and DAL (Ours). Random Sampling is the most intuitive selection strategy. Although there is no selection standard for random strategy, it is still an indispensable comparison method to measure the performance of other active learning strategies. In addition, Random Sampling represent the non-active learning methods

Regarding the metrics, we use a common rectangle metric proposed by Jiang *et al.* [45], which is also used in the paper of GG-CNN, to evaluate the performance of the backbone grasp detection network. A valid successful grasp should satisfy the following two conditions: (a) difference between the predicted grasp angle and the ground truth grasp angle to be less than 30°, and (b) jaccard index ($J(A,B) = \frac{|A \cap B|}{|A \cup B|}$) between the two grasp rectangles to be more than 25%.

B. Result on Cornell Dataset

We take 80% of the Cornell Grasp Dataset as the training set and the remaining 20% as the test set. The initial labeled pool consists of 60 randomly selected RGB-D images. For each active learning iteration, 60 RGB-D images are chosen from the unlabeled pool and added to the labeled pool according to each active learning strategy. We perform 10 trials for each active learning strategy with the same random seed in every trial. Finally, we calculate the average accuracy of all iterations.

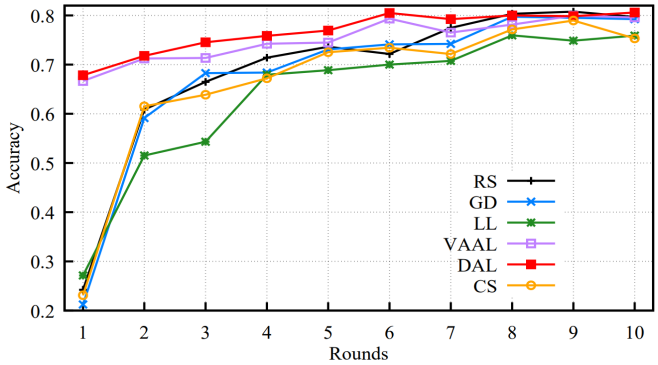


Fig. 6. Comparison on the Cornell Grasp Dataset [45] including RS (Random Sampling), GD (Graph Density [23]), LL (Learning Loss [20]), VAAL [19], CS(Core-set [22]) and DAL (Ours). The x-coordinate represents the active learning selection rounds and the y-coordinate represents the average accuracy of 10 trials.

As is shown in Fig. 6, our proposed active learning framework DAL achieves the highest performance during 10 rounds active learning selection. Although the labeled pool is quite small in the early stage, DAL takes advantage of the latent features of both labeled data and unlabeled data. Therefore, DAL shows outstanding performance with

few labeled data. In each round, DAL trains the shared encoder more sufficiently than other strategies. The grasp pose detection decoder and the active learning discriminator both benefit from the well-trained encoder. Thus, DAL is able to discriminate more representative data samples. VAAL also shows relatively high performance compared to the other strategies, however, it ignores the natural latent features provided by the backbone GG-CNN. The performance of Graph Density and Core-set is close to that of Random Sampling, while Learning Loss does not perform well on the Cornell Grasp Dataset.

C. Result on Jacquard Dataset

As mentioned above, we build a subset of the Jacquard Grasp Dataset [47] for a more realistic comparison. The subset consists of 300 objects with 1500 RGB-D images. We take one random image of each object to build a test set that contains 300 images, thus the size of the initial unlabeled pool is 1200. Then we randomly select one image of each object to constitute the initial labeled pool. For each active learning iteration, 100 images are selected from the unlabeled pool and added to the labeled pool. We perform 10 trials for each active learning strategy and the average accuracy of each iteration is shown in Fig. 7.

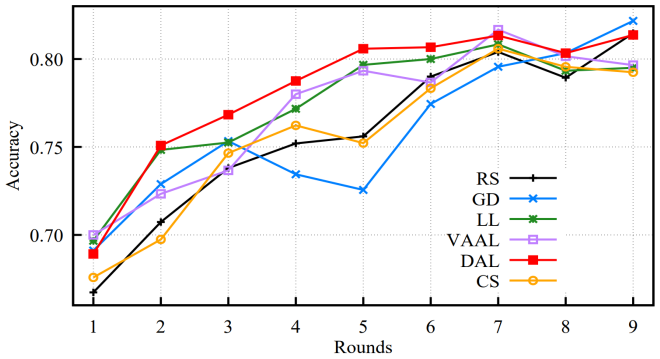


Fig. 7. Comparison on the Jacquard Grasp Dataset [47] including RS (Random Sampling), GD (Graph Density [23]), LL (Learning Loss [20]), VAAL [19], CS(Core-set [22]) and DAL (Ours). The x-coordinate represents the active learning selection rounds and the y-coordinate represents the average accuracy of 10 trials.

As illustrated in Fig. 7, our proposed active learning framework DAL outperforms the other strategies in most rounds. All strategies have similar performance in the early stage because we provide a sufficient initial labeled pool. Nevertheless, DAL still picks out informative data which helps to improve the performance of GG-CNN. Learning Loss shows better performance on the Jacquard Grasp Dataset than that on the Cornell Grasp Dataset because of the larger initial labeled pool. The performance of Graph Density declines on the Jacquard Grasp Dataset. According to our analysis, the distribution of the Jacquard Grasp Dataset may not be suitable for Graph Density to select diverse samples. As for Core-set, it performs poorly in the initial stage, but when it comes to the final round it ranks closely to DAL.

D. Result on noisy Cornell Dataset

We can not ask annotators to provide 100% accurate annotations in real life. Especially for the tasks of robotic grasping, human annotators annotate data by experience, which leads to some noise on the ground truth we used in training. To simulate real-life annotation circumstances, we apply random noise on the label of the Cornell Dataset to mock a noisy grasp dataset. In particular, we randomly add the angle of each grasp annotation with a value between $-45°$ and $45°$.

Fig. 8 illustrates the comparative result on the noisy Cornell Grasp Dataset. Although the annotation noise affects all active learning strategies to some extent, our proposed DAL still shows superior performance in most rounds. DAL pays more attention to the latent features extracted by the shared encoder, so it can discriminate which data sample is close to the labeled pool. DAL selects informative data samples that the labeled pool has no similar data with. Thus, DAL performs more stable to annotation noise. On the contrary, other strategies especially Random Sampling barely utilize the natural latent features, so their performance tends to be less stable.

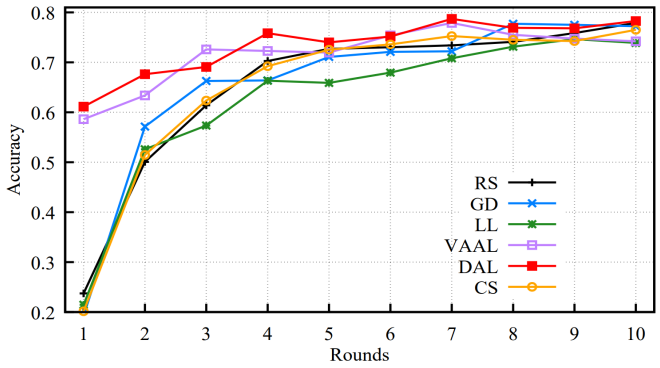
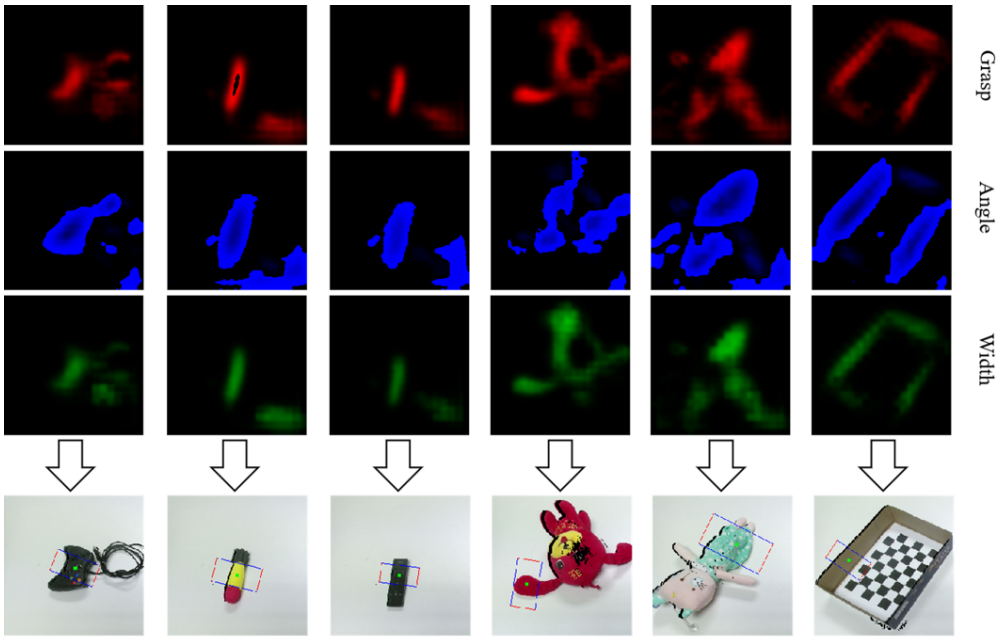


Fig. 8. Comparison on the noisy Cornell Grasp Dataset including RS (Random Sampling), GD (Graph Density [23]), LL (Learning Loss [20]), VAAL [19], CS(Core-set [22]) and DAL (Ours). The x-coordinate represents the active learning selection rounds and the y-coordinate represents the average accuracy of 10 trials.

E. Experiment on real robot

To demonstrate the validity of models trained by our proposed active learning framework, we also perform an experiment of real-time grasp pose detection on a real robot Kinova MOVO. A model trained with about 60% of the Cornell Grasp Dataset is utilized in the experiment. Several common objects different in many aspects are chosen to test the performance of the model in cluttered scene. The input images are collected by a Kinect2 RGB-D camera fixed on the top of the Kinova MOVO robot. The resolution of input images is 640 pixels$\times$480 pixels and trimmed into 300 pixels$\times$300 pixels before entering the model.

Fig. 9 shows the real-time detection results of several objects. We visualize the grasp quality map, the angle map and the width map. Besides, we draw the grasp rectangle of the best grasp on the trimmed RGB image to show qualitative results. As is shown in (a), (b), (c) of Fig.8, the model can detected the objects in Cornell Grasp Dataset well. Even for (d), (e), (f) which are unseen objects, the model also performs properly. Furthermore, we apply the model on Kinova MOVO and perform a grasp task. Fig. 5 shows the



(a) controller (b) missile (c) remoter (d) dog doll (e) cat doll (f) box

Fig. 9. Real-time detection results with the GG-CNN model trained on 60% data of the Cornell Grasp Dataset. (a), (b) and (c) are similar objects in the Cornell Grasp DataSet, while (d), (e) and (f) are unseen objects.

third-person perspective and the first-person perspective of Kinova MOVO grasping an object.

Moreover, we try to grasp each object shown in Fig. 5 for 5 times with the model trained only 4 rounds. The experiment results are shown in Table. II. Our proposed framework performs better than the state-of-art.

TABLE II

EXPERIMENT ON REAL ROBOT. OBJECT INDEX IS SHOWN IN FIG. 5

Object index	#1	#2	#3	#4	#5	#6	#7
DAL (proposed)	5/5	4/5	3/5	3/5	4/5	4/5	5/5
VAAL (comparative)	4/5	4/5	4/5	2/5	2/5	1/5	4/5

Object index	#8	#9	#10	#11	#12	Overall
DAL (proposed)	3/5	4/5	4/5	3/5	4/5	**76.7%**
VAAL (comparative)	3/5	4/5	3/5	3/5	4/5	63.3%

In the video we provide, we show the whole pipeline of our proposed active learning framework DAL and a real-world grasp experiment using the model trained by DAL.

V. DISCUSSION

Our proposed active learning framework contains about 2k parameters which is an insignificant number compared to grasp detection algorithms. During the experiments on Kinova MOVO, we test the computation time of our proposed framework DAL to generate a grasp pose and it is less than 30ms. Therefore, DAL is capable of undertaking a real-time detection task in the real environment.

As mentioned in Section III, there is a difference in the data usage during the training stage between our proposed DAL and the original GG-CNN. The shared encoder in DAL is trained with both labeled data and unlabeled data, which means it is trained more sufficiently than the encoder in the original GG-CNN. We use the data selected by DAL after 6 rounds to train the original GG-CNN and it achieves an accuracy of 77.45%, while the model trained by our proposed framework can achieve an accuracy of 80.48% using the same data. The result demonstrates that our proposed active learning framework DAL can not only select the informative data, but also help the grasp detection network to be trained better and achieve higher performance with the same data.

VI. CONCLUSION

In this paper, we proposed an active learning framework DAL for robotic grasping tasks. It utilizes a shared encoder to extract latent features of both labeled and unlabeled data. With sufficient latent features, a discriminator is established to predict a probability for each data sample. The data sample with the least probability is chosen to be annotated because it is considered as the data that is farthest from the labeled pool. Moreover, a user interface is developed in our proposed active learning framework to provide an efficient and convenient annotation tool. We demonstrate superior results on two datasets, the Cornell Grasp Dataset and the Jacquard Grasp Dataset. DAL shows better performance than other active learning strategies, especially when the size of the labeled pool is relatively small. Considering annotation noise, we build a noisy Cornell Grasp Dataset and show that our proposed method is stable to annotation noise. We also demonstrate that the model trained with about 60% of the Cornell Grasp Dataset selected by our proposed active learning framework can handle a real-world grasp detection task. In the future, we plan to generalize DAL to other deep learning applications of robotics, like navigation and scene reconstruction.

REFERENCES

[1] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proceedings 2000 ICRA. Millennium conference. IEEE international conference on robotics and automation. Symposia proceedings (Cat. No. 00CH37065)*, vol. 1. IEEE, 2000, pp. 348–353.

[2] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesisa survey," *IEEE Transactions on robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[3] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 326–336, 2012.

[4] K. B. Shimoga, "Robot grasp synthesis algorithms: A survey," *The International Journal of Robotics Research*, vol. 15, no. 3, pp. 230–266, 1996.

[5] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," *arXiv preprint arXiv:2202.03631*, 2022.

[6] 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.

[7] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 3406–3413.

[8] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *RSS*, 2018.

[9] E. Johns, S. Leutenegger, and A. J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4461–4468.

[10] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[11] E. Ogas, L. Avila, G. Larregay, and D. Moran, "A robotic grasping method using convnets," in *2019 Argentine Conference on Electronics (CAE)*. IEEE, 2019, pp. 21–26.

[12] 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.

[13] L. Antanas, P. Moreno, M. Neumann, R. P. de Figueiredo, K. Kersting, J. Santos-Victor, and L. De Raedt, "Semantic and geometric reasoning for robotic grasping: a probabilistic logic approach," *Autonomous Robots*, vol. 43, no. 6, pp. 1393–1418, 2019.

[14] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 769–776.

[15] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network. arxiv," *arXiv preprint arXiv:1909.04810*, 2019.

[16] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.

[17] D. E. Graff, E. I. Shakhnovich, and C. W. Coley, "Accelerating high-throughput virtual screening through molecular pool-based active learning," *Chemical science*, vol. 12, no. 22, pp. 7866–7881, 2021.

[18] M. Gorriz, A. Carlier, E. Faure, and X. Giro-i Nieto, "Cost-effective active learning for melanoma segmentation," *arXiv preprint arXiv:1711.09168*, 2017.

[19] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5972–5981.

[20] D. Yoo and I. S. Kweon, "Learning loss for active learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 93–102.

[21] U. G. Nagpal and D. A. Knowles, "Active learning in cnns via expected improvement maximization," *arXiv preprint arXiv:2011.14015*, 2020.

[22] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," *arXiv preprint arXiv:1708.00489*, 2017.

[23] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3626–3633.

[24] S. D. Jain and K. Grauman, "Active image segmentation propagation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2864–2873.

[25] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The power of ensembles for active learning in image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9368–9377.

[26] C.-A. Brust, C. Käding, and J. Denzler, "Active learning for deep object detection," *arXiv preprint arXiv:1809.09875*, 2018.

[27] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer handbook of robotics*. Springer, 2016, pp. 955–988.

[28] Y. Wang, Y. Zheng, B. Gao, and D. Huang, "Double-dot network for antipodal grasp detection," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 4654–4661.

[29] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

[30] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "Regnet: Region-based grasp network for end-to-end grasp detection in point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 474–13 480.

[31] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, p. 1687814016668077, 2016.

[32] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.

[33] B. Settles, "Active learning literature survey," 2009.

[34] C. Long and G. Hua, "Multi-class multi-annotator active learning with robust gaussian process for visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2839–2847.

[35] C. Long, G. Hua, and A. Kapoor, "A joint gaussian process model for active visual recognition with expertise estimation in crowdsourcing," *International Journal of Computer Vision (IJCV)*, vol. 116, no. 2, pp. 136–160, 2016.

[36] G. Hua, C. Long, M. Yang, and Y. Gao, "Collaborative active visual recognition from crowds: A distributed ensemble approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, vol. 40, no. 3, pp. 582–594, 2018.

[37] Y. Qiao, J. Zhu, C. Long, Z. Zhang, Y. Wang, Z. Du, and X. Yang, "Cpral: Collaborative panoptic-regional active learning for semantic segmentation," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[38] M. Bilgic and L. Getoor, "Link-based active learning," in *NIPS Workshop on Analyzing Networks and Learning with Graphs*, vol. 4, 2009.

[39] Y. Guo, "Active instance sampling via matrix partition," *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[40] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 79.

[41] D. D. Lewis, "A sequential algorithm for training text classifiers: Corrigendum and additional data," in *Acm Sigir Forum*, vol. 29, no. 2. ACM New York, NY, USA, 1995, pp. 13–19.

[42] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 287–294.

[43] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2016.

[44] P. Hemmer, N. Kühl, and J. Schöffer, "Deal: deep evidential active learning for image classification," in *Deep Learning Applications, Volume 3*. Springer, 2022, pp. 171–192.

[45] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International conference on robotics and automation*. IEEE, 2011, pp. 3304–3311.

[46] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[47] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3511–3516.