2022 7th IEEE International Conference on
Advanced Robotics and Mechatronics
July 9-11, 2022, Guilin, China

2022 International Conference on Advanced Robotics and Mechatronics (ICARM) | 978-1-6654-8306-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICARM54641.2022.9959267

# Unsupervised Representation Learning for Visual Robotics Grasping

Shaochen Wang, Zhangli Zhou, Hao Wang, Zhijun Li, and Zhen Kan

*Abstract*— **Despite tremendous success achieved by deep learning in the field of robotic vision, it still requires massive amounts of manual annotations and expensive computational resources to train a high-performance grasping detection model. The difficulties (e.g., complicated object geometry, sensor noise) all pose challenges for grasping unknown objects. In this paper, self-supervised representation learning pre-training is investigated to tackle the issues like expensive data annotation and poor generalization to improve visual robotics grasping. The proposed framework has two primary characteristics: 1) Siamese networks integrated with metric learning capture commonalities between similar objects from unlabeled data in an unsupervised fashion. 2) A well-designed encoder-decoder architecture with skip-connections, fusing low-level contour information and high-level semantic information, enables a spatially precise and semantically rich representation. A key aspect of using self-supervised pre-training model is that it alleviates the burden on data annotation and accelerates model training. By fine-tuning on a small number of labeled data, our method improves the baseline which does not use deep representation learning by 9.5 points on the Cornell dataset. Our final grasping system is capable to grasp unseen objects in a variety of scenarios on a 7DoF Franka Emika Panda robot. A video is available at https://youtu.be/Xd0hhYD-IOE.**

## I. Introduction

Automatic grasping, as a fundamental skill, is essential for both industrial and domestic robots. Currently, most industrial robotic manipulators are still based on fixed position grasping to perform repetitive tasks. Perceiving from high-dimensional unstructured observations, e.g., images, still remains a challenge for robots to interact with the world and accomplish their goals autonomously. Learning is a key enabling technology that significantly improves the capacity of robot perception and decision making.

Recently, learning-based approaches have drawn more and more attention. Deep learning has revolutionized many areas, e.g., visual recognition [1], natural language processing [2]. Convolutional neural networks have become the dominant tool in the field of robotic vision [3]. Most approaches in grasp detection rely on supervised learning. However, the success of these systems depends on too much manually labeled training data and tremendous computing resources. In the field of visual grasping, this issue is further exacerbated by the immense costs involved in acquiring valid grasping annotations. Due to the uncertainties of the shape, size, and pose of objects, the robots will inevitably encounter diverse types of objects when performing various tasks. Once the shape and position of an object varies slightly, a new annotation is often required leading to a time-consuming and labor-intensive task. For instance, the authors [4] spent two months collecting over $8 \times 10^5$ grasp attempts with 6 to 14 manipulators. Lenz et.al [3] applied manual annotation for all samples in the Cornell grasp dataset. Instead, in a fully supervised training mode, inadequate training samples will affect the accuracy and robustness of the resulting model. Moreover, the generalization of the model trained with insufficient data is also limited. For example, a traffic recognition system trained for daylight may not work properly at night [5]. As a result, inadequate grasping labeled data in a supervised learning manner tends to lead a poor grasping model, which has become the main bottleneck for the application of learning-based approaches in robots.

How to utilize large amounts of unlabeled data to enhance the generalization of visual robotic grasping model has become a challenging problem to be addressed. Currently, the commonly used methods for grasping detection are to train on collected datasets or in a simulator, and then deploy the resulting model to a real robot. Compared with the real unstructured environments, the collected dataset is simpler and has less noise. However, in the real world, many factors such as the different spatial layouts, partial occlusion between objects, and variations in camera viewpoint can significantly affect the accuracy of detection. A conclusive comprehension of improving the generalization of grasping detection systems is still quite lacking in the learning-based literature. Recent research advances such as transfer learning [6], unsupervised learning [7], and domain adaptation [8], etc, provide promising insight. Inspired by cognitive science, researchers are more concerned with how the neuron represents, processes, and transforms information. Humans spend most of their time learning in an unsupervised manner. A baby learns to recognize and classify faces, not because he receives supervision or rewards, but seeing many of them. Owing to recent milestones such as BERT [2], GPT [9] in natural language processing, pre-training models combined with representation learning capture rich domain knowledge and encode them in the parameters of neural networks. Representation learning works by projecting high dimensional data to a low dimensional embedding space, making it easier to find patterns and better understand the features of the data. *Can representational learning contribute to visual robot grasping?* We employ self-supervised representation learning to learn well-formed data representations

This work was supported in part by the National Natural Science Foundation of China under Grant U2013601, and Grant 62173314.

Shaochen Wang, Zhangli Zhou, Hao Wang, Zhijun Li, and Zhen Kan are with the Department of Automation, University of Science and Technology of China, Hefei 230026, China.

Zhen Kan is the corresponding author (zkan@ustc.edu.cn).

from unlabeled samples, and then use limited label data for fine-tuning. Concretely, we use siamese networks to maximize the similarity of the same concept under different views where images with different data augmentation are fed into siamese networks to minimize their distance in the latent space.

In fact, many objects have similar components, and such similarity-based representation can be generalized to many types of items. For instance, although coins and bottle caps are of distinct types, the experience of grasping coins can be easily generalized to bottle caps. Part level area representation is equally important. As an example, many objects have grasping handles and the model obtained by self-supervised learning can transfer such knowledge to other unknown items that have similar handles in the latent space. In this paper, we focus on incorporating self-supervised representation learning into visual grasping detection. Our aim is to investigate the feasibility of encoding high-level visual representation without using labeled data, and to better understand the semantic information to promote visual grasping. Our approach is superior to the current state-of-the-art grasping detection methods.

The contribution of this paper can be summarized as follows: 1) we explore self-supervised representation learning pre-training to learn a sound representation from unlabeled data for better performing visual robotics grasping tasks. 2) we have meticulously designed the network architecture for the grasp detection task, which fuses the local and global features and enables a more precise detection. 3) we conduct extensive experiments. The results show that pre-training with self-supervised representation learning can greatly accelerate the training speed of the grasping model and improve the final accuracy by $9\%$ over the model without deep representation learning.

## II. Method

### A. Problem Formulation

Grasping detection is the process of locating an item and generating the grasping pose for that item. It is a significant part of the grasping pipeline where detected grasping boxes are usually used for subsequent high-level tasks. We consider the problem of automatically generating grasping poses. The widely used grasp representation is to express the graspable region as a rectangle proposed in [3]. A grasping configuration is parameterized by a 5 dimensional tuple defined as:

$$g = (x, y, \theta, h, w) \tag{1}$$

where $(x, y)$ is the central coordinate of the grasping rectangle in the image, and $\theta$ is the rotation angle of the gripper along the z-axis. $h$ and $w$ are the height and width of the rectangle, respectively. Since the rectangle representation does not reflect the quality of the generated grasp rectangle. Multiple candidates have to be generated and the best among them can then be picked, which is time-consuming.

To address the issue above, we use the grasp maps [10] defined as

$$\mathbf{G} = (\Theta, W, Q) \in \mathbb{R}^{3\times H\times W}, \tag{2}$$

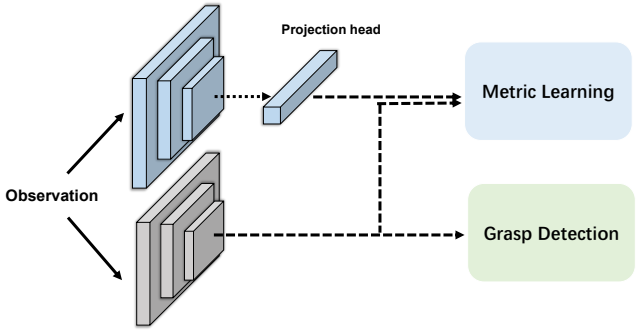



Fig. 1. Illustration of our self-supervised representation learning pipeline. The overall training is divided into two stages. Phase 1: Observations with different data augmentations are fed as positive samples into the siamese network. The distance in the latent space is minimized by metric learning between the output of the top branch encoder passed by a MLP projection and the output of the bottom branch encoder. Phase 2 is a process of fine-tuning with a small amount of label data. The output of the bottom branch encoder is passed into the decoder in the grasp detection procedure.

which is a pixel-level method to describe the grasping status. As compared to the rectangle representation, the grasp maps augment the measure of grasp quality, and are more fine-grained. $\mathbf{G}$ includes the grasp quality map $Q$, the grasp angle map $\Theta$, and the grasp width map $W$. The grasping quality map $Q \in \mathbb{R}^{H\times W}$, is a heatmap of the same size as the input picture and $H, W$ are the height and width of the input image respectively. The value of each pixel in $Q$ is a score as an indicator to evaluate the possibility of successful grasp at that point. All the pixel values in the grasping quality image are clipped between 0 and 1, where a large value indicates a high probability of successful grasping at that position. Each pixel value in the $W$ and $\Theta$ maps represents the corresponding width and angle of the gripper at that position for grasping. In contrast to the classical method of generating grasping rectangle candidates, the grasp maps approach is more time-efficient, which can directly obtain the best grasp from the quality map, and then extract the width and angle from the width map $W$ and the angle map $\Theta$.

### B. Unsupervised Visual Representation

Human can easily learn to pick up objects. For infants, when the appearance, position, etc. of objects varies, they can still easily adapt to. However, it remains a challenge for robots to cope with variations in the shape or position of objects to perform a reliable grasp. Inspired by the intuition that human can learn to extract the similarities of different concepts, learn to reason about abstract representations, and transfer the knowledge to unseen situations, we leverage self-supervision learning to make the model learn the representation that captures the potential common inherent properties either task-oriented or semantic.

In this paper, we employ siamese networks with weight sharing. The transformed images are sent to the two branches of the siamese network. For an RGB-D image $x$, two correlated views $\tilde{x_1}$ and $\tilde{x_2}$ are acquired through different data augmentation operators. Images obtained after data

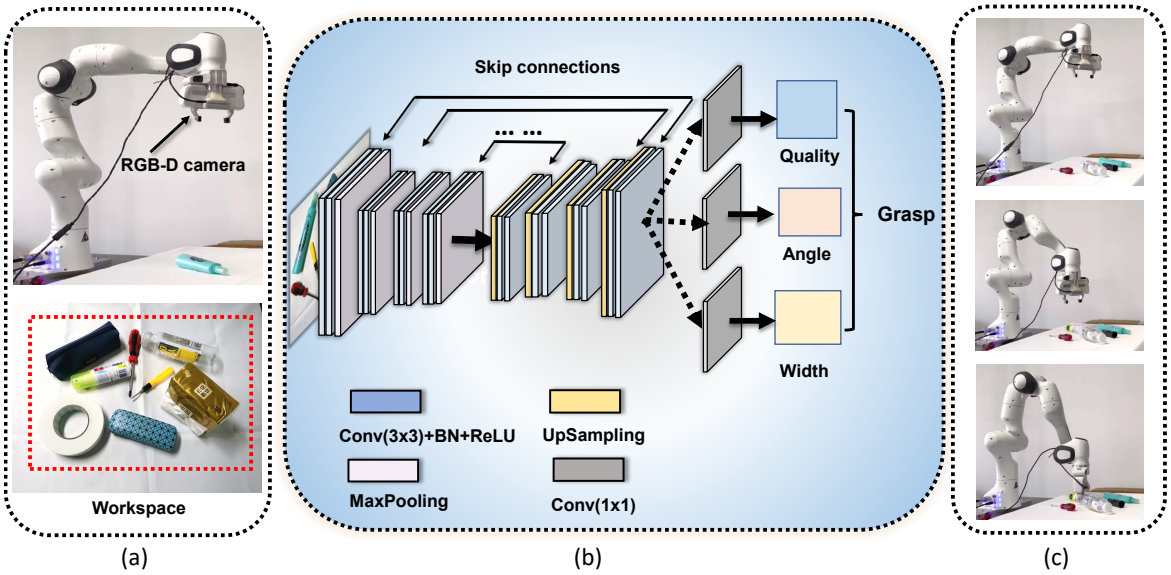



Fig. 2. (a) shows our grasping system with the RGB-D camera mounted on the top of the end-effector. And the test grasping objects are randomly placed in the workspace. (b) displays the overview of neural network architecture. Both the encoder and decoder are fully convolutional networks. (c) illustrates the process of online grasping.

augmentation are treated as positive samples of the original image. Then the augmented images are fed into the siamese encoders. And augmented samples fed into the model maximize their similarity by a metric learning loss. The encoder used in our architecture is denoted as $f$, and behind the encoder is a MLP projection head, referred to as $h$. Next, the output view of one branch transformed by the encoder and projection head, denoted as $p_1 \overset{\text{def}}{=} h(f(\tilde{x_1}))$, matches the other output vector view, expressed as $z_2 \overset{\text{def}}{=} f(\tilde{x_2})$. The distance function is expressed as the negative cosine similarity function:

$$D\left(p_1, z_2\right) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}, \tag{3}$$

where $\|\cdot\|$ means $l_2$ norm function. We utilize Eq. (5) to minimize the representation distance of the object at different views in the latent space.

The intuition is the fact that the semantic information of the same type of objects under different observations should be similar. It is desired that the model can learn this invariance through metric learning loss. In addition, we follow the training protocol used in [11] and use positive samples. Similar to [12], we adopt a symmetrized loss defined as:

$$L = D\left(p_1, z_2\right) + D\left(p_2, z_1\right), \tag{4}$$

where $D$ means the distance function mentioned in Eq. (3). $h_1, h_2$ indicates the latent representation through the encoder and projection head, while $z_1, z_2$ denotes the output vector only passing through the encoder. Moreover, to avoid a collapsing solution, a critical component is the stop gradient operator. Correspondingly, the loss function is modified as:

$$L = D\left(p_1, \text{stopgrad}(z_2)\right) + D\left(p_2, \text{stopgrad}(z_1)\right), \tag{5}$$

and in this term, $z_2$ can be regarded as a constant term after stopgrad operation and will no longer be updated by gradients. Stopgrad has been empirically validated in [12].

In the real world, there are a number of commonalities between objects, even though they may be different as a whole. For example, coins and plates belong to distinct categories of objects, but they do have similar shapes. The grasp configuration generated for plates is also beneficial for grasping coins. Likewise, even if the overall shape is not similar, learning a good representation of part area of the objects can also allow potential generalizations to many other items. For instance, many things have grasping handles, and the successful detection of the handle allows the model to transfer grasping knowledge to many unknown objects with handles. In addition, datasets used in most previous methods mainly consider objects in the center of pictures without background noise. However, in real unstructured environments, the occlusion in cluttered scenes, and different placement of objects can significantly influence the performance of the grasping. In this paper, we consider different forms of data augmentation, including geometric and spatial transformations of input images. The original image may be randomly rotated, flipped, shifted, and cropped. The siamese network aims to extract invariant features through metric learning loss to allow the robots to capture invariant representations of object pose from different viewpoints. The metric learning loss maintains an effective representation in the latent space by measuring the similarity among correlated views of the same objects. The intuition behind our approach is to leverage the self-supervised tasks to achieve "free" annotation and then use the resulting model to make valid predictions in real scenarios.

During the real robot grasping phase, we no longer need

to perform data augmentation and projection head. Instead, directly put the RGB-D image captured by the camera into the our model and the neural network outputs the corresponding grasp pose.

### C. Neural Network Architecture

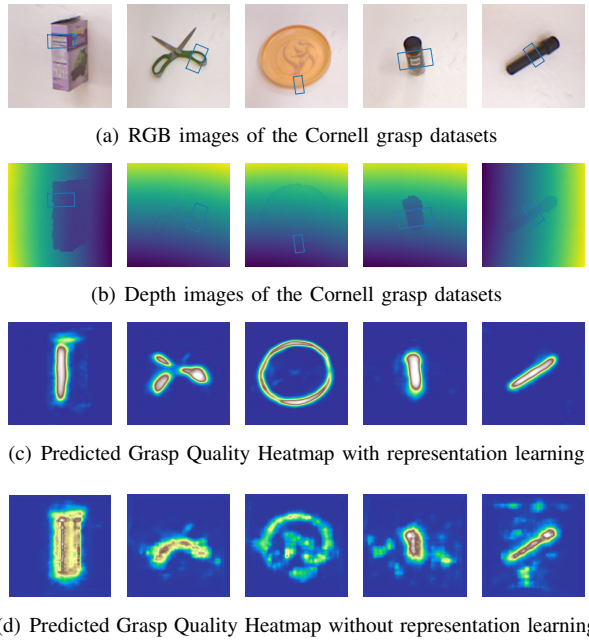
For a robotic grasping system, upon the robot observing an object, the robotic system needs to generate grasping poses (grasping coordinates and orientation angles), and moves its gripper. Due to factors such as the complex geometry of objects in the real world, it is essential for the model to acquire a spatially precise representation of these objects and robust to different positions. Inspired by recent work [13], we adopt a fully convolutional network to avoid the massive number of parameters brought by fully connected layers. And the overall network model is an encoder-decoder architecture, in which the encoder is composed of several convolution layers that gradually decrease the spatial size of feature maps. The input image is mapped to a latent space, which is formed to abstract from a high resolution to a low compact resolution representation. And the decoder consists of a deconvolution neural network which decodes the latent code into a grasping quality heatmap of the same size as the input size. In our network structure, each convolution layer is followed by a batch normalization layer [14] and then through a nonlinear activation (ReLU) transformation. The entire framework is flexible and easy to implement. Moreover, to facilitate the flow of information in the network, we apply skip connections which skip several layers in the neural network and use the outputs of one layer as inputs to the subsequent layer. Inspired by [15] [1], our model can better fuse global information (e.g. shape of the objects) and local information such as detailed texture of items. The network takes in RGB-D images as input and generates three pixel-level grasping quality, angle, and width heatmaps. The resulting grasping point is the location with the highest quality score in the grasping quality map. By introducing self-supervised learning, the encoder learns a compact representation, filtering out redundant information from input data. Meanwhile, through a variety of data augmentation techniques, the representation is insensitive to the rotation and shift of objects with the help of metric learning loss. An optimal grasp candidate is inferred from both the grasping quality map, and the generating grasping pose used for subsequent grasp planning. The loss is defined as the mean square error, which compares the average difference between predicted value and ground truth,

$$L = \|G - \hat{G}\|^2 \tag{6}$$

where $\hat{G}$ is the ground- truth, and $G = (\Theta, W, Q)$ is the three heatmaps of the network output.

## III. Experiments

In this section, we conduct several experiments to demonstrate how the method described in the previous section can be used for visual robotics grasping. Our main focus lies in two folds. The first concern is the quality of our representation learning approach. Is it helpful to boost the accuracy of grasping detection? We compare our approach with the state-of-the-art methods on Cornell grasping dataset. Besides, we also visualize the heat map learned by the model with our self-supervised representation learning. The second focus is whether the visual representation learning facilitates the generalization of grasping models on real robots. Moreover, in the real world, the noise, object friction, etc in visual perception all affect the predicted grasp scaling to real robots.



(a) RGB images of the Cornell grasp datasets

(b) Depth images of the Cornell grasp datasets

(c) Predicted Grasp Quality Heatmap with representation learning

(d) Predicted Grasp Quality Heatmap without representation learning

Fig. 3. A comparison of the grasp quality heatmap between methods with and without self-supervised representation learning on Cornell grasp dataset. The first and second rows show the RGB and depth images of the items. The rectangles displayed in the pictures are the predicted grasping rectangles by our method. Each pixel in the grasp quality heatmap represents the probability of successful grasping when grasping at this position, where a brighter color means a greater grasp success probability. On the contrary, the darker the color, the lower the probability.

### A. Dataset and Experiment Configuration

The Cornell grasping dataset is a multi-object dataset which consists of 885 RGB-D images of 240 indoor scene objects. Each object in the dataset is provided with multiple human annotations for grasping. The ground-truth is presented in the form of grasp-rectangles around the objects. The data is randomly split into training set and test set. In terms of the accuracy evaluation metric, we utilize the method proposed in [16]. An effective grasp needs to meet the following rules: 1) the predicted grasp angle shows a difference of less than $30^\circ$ compared with the ground-truth 2) the Jaccard index calculated from the predicted grasp and the ground truth is greater than $25\%$. The Jaccard is defined as: $J\left(R^*, R\right) = \frac{|R^* \cap R|}{|R^* \cup R|}$, where $R^*$ refers to the ground truth and $R$ represents the predicted grasp. Indeed, the Jaccard index is an indicator that measures how well the predicted grasp matches the human labels.

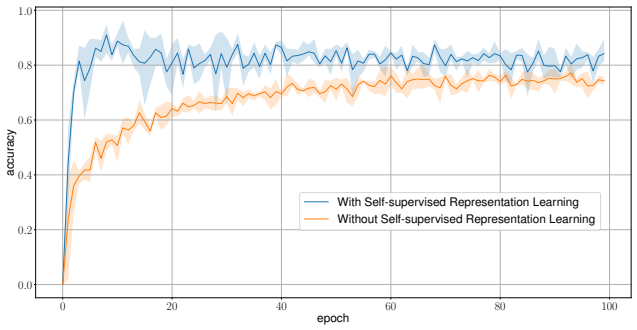



Fig. 4. The accuracy of the approaches with and without self-supervised representation learning on the test set in the Cornell grasp dataset during training phase. We can observe that applying self-supervised representation learning can accelerate the training speed of the model.

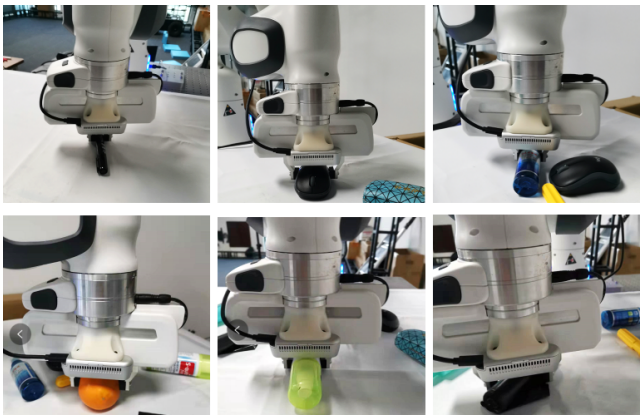


Fig. 5. Visual grasping results on unseen objects.

The architecture of our neural network is depicted in Fig. 2. The encoder consists of four blocks, which in turn contains $\{64, 128, 256, 512\}$ feature channels. Each block contains convolutions ($3 \times 3$ kernels), ReLU activation, BN layers, and after that, a maxpooling with stride 2. The decoder is composed of upsampling layer that gradually restores the output to the same size as the input dimension. The last layer is a $1 \times 1$ convolution to adjust the number of channels. Both the encoder and decoder are optimized by using the Adam [17] optimizer with a learning rate of 1e-4. And the mini-batch size is 32. The original image size is $640 \times 480$. To reduce the computational cost, the image is resized to $300 \times 300$. The model is trained by standard back propagation and the entire training procedure will be completed within 100 epochs. The skip connection is implemented by concatenation with feature channels.

### *B. Analysis*

Many objects that do not appear in the dataset are selected as grasp items to test the performance of the model in the real robot system. Each object is grasped several times with a random placing position and orientation. Fig. 5 shows the grasping process of our approach. The camera is placed on the robot, which is parallel to the desktop. The items are positioned in the field of camera view. The image captured by the camera is cropped to a square size of $300 \times 300$ after image processing. Immediately, the image is fed into the neural network to generate a pixel-level grasp quality map, and the best grasp pose is calculated. Next, the robot moves down its end-effector until the pre-calculated grasp pose is met. During this process, once a collision is detected, it is considered as a failure grasp. Furthermore, the grasping objects have different shapes and geometry, which requires a high precision of the model. The grasping results on unseen objects illustrate that the neural network with representation learning can predict robust poses. We also conduct experiments with multiple objects grasp. Despite the fact that our model is trained on a single object, it can be well adapted to multi-object environment with the help of self-supervised learning. In cluttered environments, it is crucial to characterize the shape of objects.

Fig. 3 shows that the visualization results of the grasping quality heatmap and grasping angle heatmap with and without self-supervised representation method, respectively. It is easy to see that the heatmap of our model is more spatially precise. We can observe that our model can readily learn areas that are easy for grasping, such as the edges of objects through self-supervised representation. Meanwhile, the model focuses on more general characteristics rather than just individual features. On the contrary, the grasp quality map learned by the model without representation learning is rather coarse and sometimes even the background area is considered to be graspable. For example, for the box shown in Fig. 3, our method evaluates the edge of the box with a higher grasp quality, while the model without self-supervised learning also identifies the flat area of the box surface as high grasp quality region. In addition, on the real robotic manipulator, the model without self-supervised pre-training is prone to collisions due to inaccurate grasps. It helps the model learn a refined representation, ignoring the noise in the input and improving the stability. Fig. 4 shows the accuracy of the model on the Cornell dataset with and without using self-supervised representation learning, where the x-axis represents the training epoch, and the y-axis means the accuracy. The experiments are conducted with 3 different random seeds where the solid line denotes the mean of the accuracy and shaded area indicated the variance.

### *C. Grasp System*

The hardware platform consists of an Intel RealSense camera D435 and a 7-DoF Franka Robot. The camera equipped with depth sensors and RGB sensors is mounted on the robot arm and the entire experimental workspace is shown in Fig. 5. In each grasping process, the captured RGB-D image by the camera is transmitted to the robot by ROS interface and the controller performs the planning and executes the generated grasping configuration on the object through the end-effector. A successful grasp is achieved by lifting the object and placing it in the specified place. All experiments are implemented on a Desktop with a 20-core Intel i9 CPU and an NVIDIA 3090 GPU.

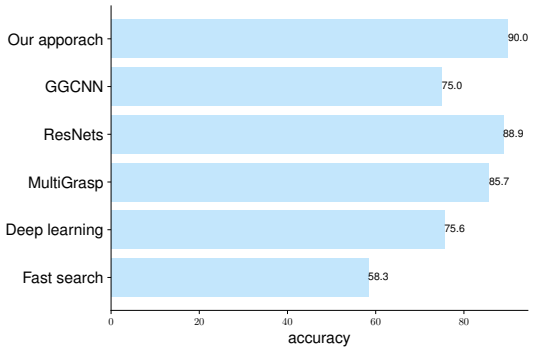



Fig. 6. The detection accuracy of different approaches on the Cornell grasping dataset.

In our robotic system, the angular rotation of the gripper is a scalar, in the interval of $[\frac{-\pi}{2}, \frac{\pi}{2}]$. Likewise, the width of gripper is in the range $[0, w_{max}]$, and $w_{\max}$ is the maximum width of gripper. Afterwards, the control module plans a trajectory with a planner and uses the generated grasping pose to perform an effective grasp. The overall control system is implemented by ROS. Although, our model performs well in most scenarios, there still exist some failure cases. For some small and thin items, it might fail to grasp them. In clutter environment, the gripper is often easy to collide with other objects, resulting in a failure grasping.

Additionally, in Fig. 6, we present the grasping detection results of our method and other state-of-the-art approaches: the first Deep Learning method [3], Grasp-ResNets [18], GGCNN [10], MultiGrasp [19], and Fast search [16]. On the whole, our method achieves an accuracy of $90\%$, outperforming other approaches. In particular, 1) compared to Deep Learning approach [3], our network achieves an improvement of 14.4 points. And our model is more effective and flexible, eliminating the need to generate multiple grasping candidates. 2) compared to prior best performed Grasp-ResNets [18], our method improves the accuracy by 1.1 points. Also, different from the Grasp-ResNets with a 50 layers of residual blocks and a 1024 nodes fully connected layer, our network is a lightweight network and has been validated on a real robot arm with good performance. 3) compared to GGCNN [10], our method still outperforms by a large margin and learns the characterized representation from unlabeled data. Moreover, in contrast to Grasp-ResNets and MultiGrasp, we also perform grasping on real robots.

## IV. Conclusion

In this work, we present a novel approach for visual grasping detection, which learns a representation that is invariant to variations in the position, shape, etc. of objects. The proposed method exploits a large amount of unlabelled data and utilizes the siamese network to make the distance of similar objects in the latent space as close as possible. Different from the currently popular methods, our method alleviates the need for annotation data and is simple to implement. More importantly, our model offers a remarkable improvement compared to the baselines methods. Our results provide insights into how to bridge merits of existing self-supervised techniques with robotic grasping. In the future, it would be interesting to investigate the integration of self-supervision with environmental uncertainty for robotic grasping.

## References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

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

[3] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robotics Res.*, vol. 34, no. 4-5, pp. 705–724, 2015.

[4] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robotics Res.*, vol. 37, no. 4-5, pp. 421–436, 2018.

[5] D. Dai and L. V. Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," *Int. Conf. Intell. Transp. Syst. Maui, HI, USA, November 4-7*. 2018, pp. 3819–3824.

[6] T. Weng, A. Pallankize, Y. Tang, O. Kroemer, and D. Held, "Multi-modal transfer learning for grasping transparent and specular objects," *IEEE Robotics Autom. Lett.*, vol. 5, no. 3, pp. 3791–3798, 2020.

[7] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434*,2015.

[8] J. Truong, S. Chernova, and D. Batra, "Bi-directional domain adaptation for sim2real transfer of embodied navigation agents," *IEEE Robotics Autom. Lett.*, vol. 6, no. 2, pp. 2634–2641, 2021.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[10] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robotics Res.*, vol. 39, no. 2-3, pp. 183–201, 2020.

[11] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv preprint arXiv:2006.07733*, 2020.

[12] X. Chen and K. He, "Exploring simple siamese representation learning," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15 750–15 758.

[13] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017.

[14] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *Int. Conf. Mach. Learn.* 2015, pp. 448–456.

[15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[16] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," *IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3304–3311.

[17] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[18] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 769–776.

[19] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *IEEE Int. Conf. Robot.Automat.*, 2015, pp. 1316–1322.