2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GraspContrast: Self-supervised Contrastive Learning with False Negative Elimination for 6-DoF Grasp Detection

Wenshuo Wang[1], Haiyue Zhu[2†] and Marcelo H. Ang Jr[3]

***Abstract*— Robotic manipulation is a grand domain that primarily involves the use of robotic arms to interact with objects in the environment. While proposed methods have achieved advancements in grasping objects, they rely heavily on extensive training data that presents a significant challenge due to the labor-intensive process of human annotation. To address the issue, we propose GraspContrast, a self-supervised contrastive learning framework leveraging unlabeled RGB-D images to enhance point-wise feature representations for 6-DoF grasp detection. Our method designs a dual-branch network architecture to learn transformations that embed positive point pairs nearby, while pushing negative point pairs far apart. Specifically, we discuss a false negative elimination strategy to explicitly detect and remove the false negative samples that undesirably repel the point instances from the geometrically similar samples. Our method exhibits consistent improvements over existing learning-based grasp detection methods on both the GraspNet-1B benchmark and physical UR10e platform. These significant performance gains demonstrate the effectiveness of our proposed framework.**

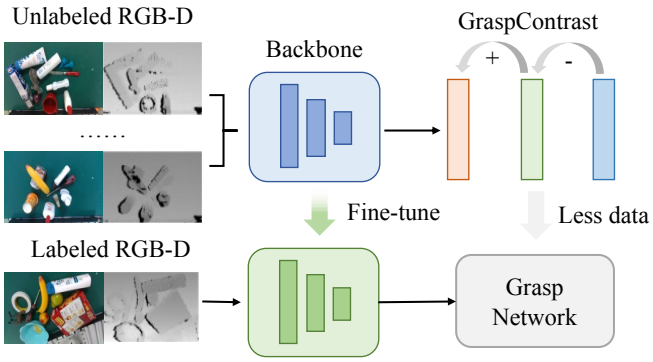

Fig. 1: Brief overview of our framework. The backbone network is pre-trained by GraspContrast to align point features of unlabeled inputs. It is then fine-tuned within the grasp detection pipeline to reduce the demand of grasp annotations.

I. Introduction

Robotic grasping has been a research challenge with both scientific and practical implications for industrial manufacturing. Recently, deep learning-based methods [1, 2, 3] have made considerable progress to enhance robotic vision and grasping capabilities. In pursuit of increasingly competitive performance, more and more complex network architectures [4, 5, 6] are explored to train on various datasets in the supervised learning manner. While many large-scale datasets (e.g. GraspNet-1Billion [7] and MetaGraspNet [8]) are collected to improve the model's generalization ability, annotating these grasping labels would be tedious and time consuming. Additionally, the proposed high architectural models concurrently increase computational complexity, thereby making their deployment on real-world robotic platforms a challenging endeavor.

To alleviate the pressure of annotating grasp labels and deploying large-scale neural networks, an intriguing avenue of research pertains to the exploration of leveraging unlabelled RGB-D images to enhance the feature representation of grasp detection models. Thanks to the rapid developments of self-supervised learning [9, 10, 11, 12] in computer vision field, unlabeled data are greatly leveraged to pre-train the neural networks and improve the model performance in downstream tasks. In the context of 3D domain, PointContrast [13] first introduces self-supervised strategy in point cloud learning and pre-trains the backbone network for scene understanding. Motivated by PointContrast in 3D representation learning, this study aims to enhance the model's performance through utilizing unlabeled RGB-D images to initialize optimal weights for the backbone network and then fine-tune it in the 6-DoF grasp detection pipeline.

Notably, most state-of-the-art methods of self-supervised learning are converging around and fueled by the central concept of contrastive learning [9]. Contrastive learning provides an effective method to learn feature representations and then leverage the information in the downstream applications. In the field of our robotic grasping task, learning effective representations plays a critical role in enhancing the capabilities of robotic systems in manipulating objects in the physical world. It allows robots to extract relevant features from sensor data, such as depth images or point clouds, which are crucial for understanding the semantic, geometrical and spatial information of objects in the environment. These representations can capture discriminative features that help in identifying graspable regions and object shapes.

Considering the potential benefits of contrastive learning in robotic manipulation tasks, we introduce a plug-and-play self-supervised learning framework named GraspContrast for

This research is supported by A*STAR "RIE2025 IAF-PP Advanced ROS2-native Platform Technologies for Cross sectorial Robotics Adoption (M21K1a0104)" programme.

[1]Wenshuo Wang is with the Department of Mechanical Engineering, National University of Singapore, 117575, Singapore wenshuo_wang@u.nus.edu

[2]Haiyue Zhu is with the Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), 138634, Singapore zhu_haiyue@simtech.a-star.edu.sg

[3]Marcelo H. Ang Jr is with Advanced Robotics Centre at National University of Singapore, Singapore 117608, Singapore mpeangh@nus.edu.sg

† Corresponding author: zhu_haiyue@simtech.a-star.edu.sg

the 6-DoF grasp detection task. The primary goal of our approach is to induce prior knowledge of the 3D space by encouraging the model to focus on the underlying structure or semantics of the point cloud data. The brief system is illustrated in Fig. 1, it mainly consists of two stages: 1) the backbone pre-training stage and 2) the grasp fine-tuning stage. In the pre-training process, we propose a dual-branch network architecture capable of simultaneously encoding features by considering unlabeled 3D point clouds from different viewpoints. Our method utilizes a pair of scene viewpoints and compute the geometric transformation between point clouds to maintain the correct correspondence between points pairs. Positive and negative sample pairs are then obtained from the correspondence mapping. In order to lead to more effective and efficient representation learning, the False Negative Elimination strategy is proposed to reduce the adverse effects of neighboring negatives during the feature representation learning. The whole network is pre-trained by our proposed contrastive learning objective that enforces the consistency between pairwise features. In the second stage, our framework builds upon the grasp detection pipeline GSNet [14]. Leveraging the pre-trained weights learned by GraspContrast, it enhances the robustness of grasp detection in the downstream task.

We conduct experiments on the public GraspNet-1Billion benchmark [7] and on a real-world robotic UR10e platform. In our experiments, we observe that the grasp detection model exhibits superior performance when fine-tuning the backbone compared to training it from scratch. Remarkably, with the limited amount of training data ($\sim$1.5%), the model achieves a 100% performance increase. Additionally, we provide extensive ablation studies to demonstrate the effectiveness of false negative elimination and data augmentation methods in our proposed method.

Our contributions can be summarized as follows:

- We propose GraspContrast, a self-supervised contrastive learning framework that leverages unlabeled images to pre-train the backbone for 6-DoF grasp detection.
- We introduce the false negative elimination strategy to increase feature diversity and alleviate the impact of false negative samples during the pre-training phase.
- The grasp detection model achieves highly competitive results with backbone fine-tuning in both the GraspNet-1Billion benchmark and real-world settings.

II. Related Works

A. 6-DoF grasp detection

The learning-based methods of 6-DoF grasp detection have attracted significant attention in the research field of robotic manipulation. GPD [15] and PointNetGPD [16] focus on grasp quality evaluation models with CNN-based or PointNet-based [17] network architecture. 6-DoF GraspNet [18] samples a variaty of grasps leveraging variational autoencoder and introduces an evaluator network to assess and refine the sampled grasps. Recently, Fang et al. [7] proposes GraspNet-1Billion benchmark, based on which many works [14, 19, 20] focus on exploring different 6-DoF grasping pipelines. Notably, GSNet [14] is the current state of the art approach that incorporates graspness model to first determine grasp centers and then predict the grasp orientations. However, almost all grasping pipelines are trained from scratch and the role of 3D representation learning in the pre-training stage has not been widely explored. Therefore, in this work, we primarily incorporate unlabeled data to first pre-train the model and then fine-tune it to improve the performance of a 6-DoF grasp detection pipeline.

B. Self-supervised learning

As a subset of unsupervised learning, self-supervised learning methods seek to learn meaningful representations from large-scale data without the need for manual annotation. According to the classification criteria in [21] to design different pretext tasks, existing methods of URL can be grouped into generative-based [22], context-based [23], cross modal-based [24], and semantic label-based categories [25]. Our work is similar to methods in the context-based category, which aims to design discriminative pre-text tasks for representation learning. Our method takes inspiration from the concept of self-supervised feature learning and explores its transferability to the 6-DoF grasp detection task.

C. Contrastive learning

As a key component of self-supervised learning, contrastive learning is aimed to encode the features of similar samples close while push the features of different samples away. Previous works [26, 27, 28] focus on instance-level approach on image classification and achieve competitive results compared to the fully supervised methods. Besides the classical 2D tasks, PointContrast [13] propose the PointInfoNCE objective to pre-train the backbone and generalize the learned representations to 3D scene understanding benchmarks. Hou et al. [29] further introduces spatial contexts into 3D pre-training, resulting in improved performance of data-efficient learning. Motivated by PointContrast, we propose GraspContrast in this work to learn effective point-wise representations for the downstream grasp pose detection task.

III. Methodology

A. System Overview

In this work, we concentrate on contrastive learning leveraging unlabeled RGB-D images to enhance feature representation for the 6-DoF grasp detection task. Our approach consists of two stages: 1) the backbone pre-training stage and 2) the grasp fine-tuning stage. During the pre-training phase, we introduce the framework, as illustrated in Fig. 2, to learn feature representations from different viewpoints inputs. The framework encodes point clouds with data augmentations into point-wise features and then apply contrastive learning to align their features. In the fine-tuning phase, we select GSNet [14] as the grasp pipeline and the initial weights of the backbone are followed by the pre-trained network. Following the pre-training procedure, the key components of our method can be divided into three parts: 1) Positive

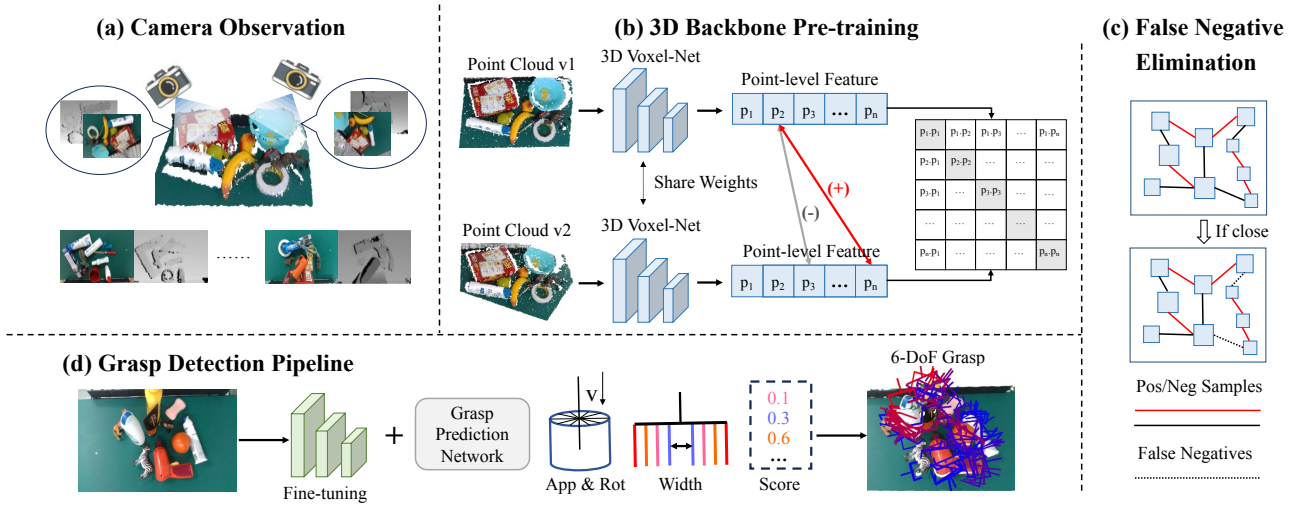

Fig. 2: Overview of our framework. (**a**) **Camera Observation.** Unlabeled RGB-D images from various scenes are captured by a camera from multiple viewpoints. (**b**) **3D Backbone Pre-training.** The 3D Voxel-Nets encode the augmented multiple-view point clouds into point-wise features, which are then aligned using the GraspContrast objective. (**c**) **False Negative Elimination.** False negative pairs are filtered out to enhance feature learning. (**d**) **Grasp Detection Pipeline.** The backbone 3D Voxel-Net with pre-trained weights are initialized and fine-tuned in the grasp detection pipeline.

and Negative Samples, 2) GraspContrast Objective and 3) Downstream Task Learning.

B. Positive and Negative Samples

The key aspect of GraspContrast lies in pre-training the backbone network by aligning feature pairs in a contrastive manner. This involves embedding positive sample pairs close to each other while trying to push negative pairs far apart. To provide a comprehensive understanding, we'll delve into the process of deriving positive and negative sample pairs.

We utilize double-view RGB-D images as inputs to establish positive and negative sample pairs. Our main objective in constructing a contrastive procedure is to generate mapping relationships between these samples. Given a cluttered scene S and a robotic arm with an eye-in-hand camera, RGB-D images of two sampled viewpoints are captured, and the depth images are projected into scene point clouds P_1 and P_2 by corresponding camera intrinsic parameters. We then apply data augmentations T_1 and T_2 to obtain the transformed point clouds $P_1^T = T_1(P_1)$ and $P_2^T = T_2(P_2)$ with additional variability and perturbation. These geometric transformations compel the neural network to learn feature equivariance by adapting to the imposed disturbances.

We consider point pair (p_1^i, p_2^i) where $p_1^i \in P_1$, $p_2^i \in P_2$ as positive if their spatial positions are close enough in the same Euclidean space. We do not use additional samples but rather the point p_1^i with its non-matched one are defined as the negative pair. Suppose we have access to key points set $p_1 = \{p_{t1}^i \in P_1^T | i = 1, \ldots, M\}$ through a certain sampling strategy, we aimed to find matched points $p_2 = \{p_{t2}^i \in P_2^T | i = 1, \ldots, M\}$ by aligning their positions across their respective viewpoints and geometric transformations. Firstly, the mapping transforming points' positions from the camera coordinate to the world coordinate is denoted as

$$(p_{w1}^i, p_{w2}^i) = (Cam_1 T_1^{-1} p_1^i, Cam_2 T_2^{-1} p_2^i), \tag{1}$$

where p_{w1}^i and p_{w2}^i represent the corresponding positions of p_1^i and p_2^i in the world coordinate. T_1^{-1} and T_2^{-1} denote the reverse transformations that transform the point clouds back into its original state. Cam_1 and Cam_2 are the extrinsic parameters of the calibrated camera. To discriminate if p_1 and p_2 are matched, we propose a criterion that measures the Euclidean distance between p_{w1}^i and p_{w2}^i, i.e.,

$$||p_{w1}^i - p_{w2}^i||_2 < \epsilon. \tag{2}$$

Due to the sensor noise, we do not apply a strict equality constraint but set a distance threshold ϵ to measure if p_{w1}^i and p_{w2}^i are close enough. Therefore, we can shortlist all matched points and obtain the positive set $\mathcal{P}^+ = \{(p_1^i, p_2^i) | i = 1, \ldots, M\}$ and negative set $\mathcal{P}^- = \{(p_1^i, p_2^j) | i, j = 1, \ldots, M; i \neq j\}$.

C. GraspContrast Objective

In the context of 3D domain, PointContrast [13] proposes PointInfoNCE loss for 3D point cloud understanding. This approach considers contrastive learning as a dictionary look-up in a softmax-based classification problem. The PointInfoNCE loss is denoted as

$$\mathcal{L}_c = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp(\mathbf{f}_i \cdot \mathbf{f}_j/\tau)}{\sum_{(\cdot,k)\in\mathcal{P}} \exp(\mathbf{f}_i \cdot \mathbf{f}_k/\tau)}, \tag{3}$$

where $\mathcal{P}$ represents point pairs across both point clouds. For a key point p_i and its correspondingly encoded feature $\mathbf{f}_i$, $\mathbf{f}_j$ and $\mathbf{f}_k$ denote the positive and negative features respectively. Also, τ is the temperature parameter for shaping the similarity scores between feature pairs. The design of this contrasting between two transformed point clouds helps the

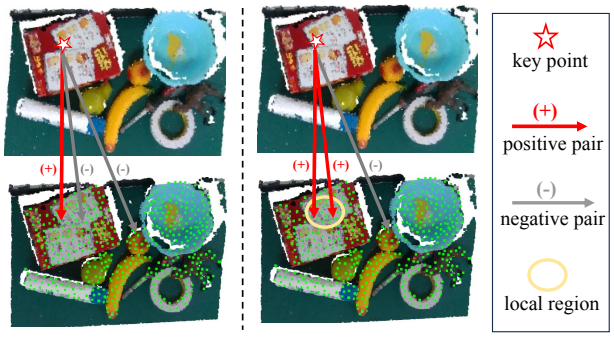

Fig. 3: Illustration of FNE strategy. (**Left**) Examples of sampled positive and negative pairs. (**Right**) The false negative samples within a similar geometrical region are filtered out by the FNE strategy.

network learn an effective 3D feature representations in an unsupervised manner and then well generalized to different downstream tasks.

While PointInfoNCE can guide basic contrastive learning in pre-training the backbone network, it still has several drawbacks. For instance, it is undesirable to push away the embeddings of false negative pairs originating from the same local region due to their apparent geometrical similarity. In order to address this deficiency, we introduce a guided sampling strategy (False negative elimination (FNE)) aimed at filtering out false negative pairs to enhance feature learning.

As illustrated in Fig. 3, some negative point pairs in $\mathcal{P}^-$ may share similar internal properties but are incorrectly regarded as dissimilar. Therefore, we denote an indicator function $NF_{i,k}$ to filter out these false negative samples for effective feature learning. The definition of $NF_{i,k}$ is based on the assumptions that a set of neighboring points preserves similar properties. Accordingly, $NF_{i,k}$ is denoted as

$$NF_{i,k} = \mathbb{1}(D_{i,k} \geq \gamma_D), \tag{4}$$

where $\mathbb{1}(\cdot)$ is an oracle function that determines whether the following condition is satisfied. $D_{i,k}$ represents the Euclidean distance of two points (p_i, p_k) in a negative point pair. γ_D is a threshold to measure if p_i and p_k are neighbors in the local region. By incorporating $NF_{i,k}$ into GraspContrast, the loss computation of neighboring sample pairs are eliminated to prevent the adverse learning process.

Our GraspContrast objective $\mathcal{L}_{gc}$ further considers this strategy based on PointInfoNCE. The formulation of $\mathcal{L}_{gc}$ is finally denoted as

$$\mathcal{L}_{gc} = -\sum_{(i,j)\in\mathcal{P}} \log \frac{\exp\left(\mathbf{f}_i \cdot \mathbf{f}_j/\tau\right)}{\sum_{(\cdot,k)\in\mathcal{P}} NF_{i,k} \cdot \exp\left(\mathbf{f}_i \cdot \mathbf{f}_k/\tau\right)}, \tag{5}$$

where $(\mathbf{f}_i, \mathbf{f}_j)$, $(\mathbf{f}_i, \mathbf{f}_k)$ indicates positive and negative feature pairs sampled by FPS. Guided by $\mathcal{L}_{gc}$, we can pre-train our backbone network.

D. Downstream Task Learning

In learning process of our downstream task, 6-DoF grasp detection, we leverage GSNet [14] pipeline, which achieves the state-of-the-art performance on GraspNet-1Billion benchmark [7]. For the network architecture, the backbone network first extracts point-level features for a scene observation. Cascaded graspness model and grasp operation model are learned to predict point-wise graspable landscape, view-wise graspable landscape, grasp scores and gripper widths respectively. We refer readers to the original paper [14] for more details. In our downstream task learning settings, we fine-tune the backbone network of GSNet in a supervised manner and the initial weights are determined by the pre-trained 3D Voxel-Net.

IV. Experiments

A. Dataset

We evaluate the performance of our framework GraspContrast on GraspNet-1Billion [7]. It is a large-scale dataset containing over 1 billion grasping annotations in real-world settings. This benchmark dataset provides overall 100 training scenes and 90 testing scenes consisting of three categories such as test seen, test similar and test novel. For each scene, 256 RGB-D images are captured by the Realsense or Kinect camera from different viewpoints. Note that the entire dataset is utilized during the pre-training phase. In the fine-tuning stage, we use different proportions of training set, e.g., 1, 16, and 256 frames for training, and the full testing set for evaluation.

B. Implementation Details

1) Pre-processing details: The raw scene point clouds are downsampled to 20000 points and then voxelized with the size of 5mm. The Farthest Point Sampling (FPS) technique is used to select a subset as key points. Additionally, we adopt random flip along global YZ plane and random rotation along upper Z-axis as point cloud augmentation methods. Note that the 3D transformations further generate the diverse variations of point clouds data captured by camera from alternate viewpoints. Augmentation methods are applied independently to facilitate precise correspondence identification between them.

2) Network details: In the pre-training stage, we leverage a dual-branch network architecture with ResUnet14 backbone built upon MinkowskiEngine [30]. Given a point cloud as input, it can generate a point-wise feature with channel C=512. In the downstream procedure, we apply the pipeline GSNet [14] as the grasp detection network. The point-wise graspable landscape, view-wise graspable landscape, grasp scores and gripper widths are predicted respectively. We refer readers to the original paper [14] for more details.

3) Training details: In the pre-training stage, the learning rate is initialized to 0.001 and decayed with a constant ratio of 0.95. We train the network using one NVIDIA 3090 GPU with Adam optimizer [31] and batch size 4. The temperature τ in GraspCpntrast is set to 0.07 and the weights of all loss terms is set to 1. In the downstream task learning, we apply

TABLE I: Evaluation results on GraspNet-1B RealSense/Kinect with different proportions of labeled frames.

Methods	1 / 256 (0.39%)			16 / 256 (6.25%)			256 / 256 (100%)		
	AP seen	AP similar	AP novel	AP seen	AP similar	AP novel	AP seen	AP similar	AP novel
GSNet [14]	3.96/2.76	3.13/2.77	1.80/1.10	37.72/26.20	34.69/25.17	15.78/10.04	65.70/61.19	53.75/47.39	23.98/19.01
Ours	**7.84/4.26**	**7.11/4.46**	**3.75/2.99**	**39.51/27.53**	**37.84/26.80**	**17.67/12.09**	**67.22/62.31**	**55.91/48.03**	**25.79/20.14**

the same training details as GSNet [14] for a fair comparison between fine-tuning from pre-trained bakcbone and training from scratch of the network.

C. Main Results

In this section, we demonstrate the effectiveness of our proposed approach on GraspNet-1B benchmark [7] based on the grasp detection pipeline - GSNet [14].

Table I illustrates the main results of our models when fine-tuning with 1/256 (0.39%), 16/256 (6.25%) and 256/256 (100%) labels. All models for both RealSense and Kinect datasets are trained separately. For Realsense with limited data, we observe that our method achieves 7.84/39.51, 7.11/37.84 and 3.75/17.67 AP metrics [7] across all test categories with 1/256 and 16/256 training frames. Compared with training GSNet from scratch, our method demonstrates improvements of 3.88/1.79, 3.98/3.15 and 1.95/1.89 AP. These performance enhancements underscore the effectiveness of GraspContrast in data-limited scenarios.

Besides, we also conduct experiments on the full training set. As shown in the last three columns of Table I, our proposed GraspContrast can still boost the performance (+1.52 AP seen, +2.16 AP similar and +1.81 AP novel). The performance gap between pre-training and training from scratch becomes larger when the data size is smaller, suggesting that self-supervised pre-training methods benefit more when the amount of labeled data is limited. This conclusion is also line with those observed in other works [32, 33] for different downstream supervised tasks. Under the condition of fine-tuning with 0.39% labels, the improvement in performance can even approach 100%. Furthermore, the performance improvements are consistent with those observed on the Kinect dataset. All results demonstrate the generalization ability of GraspContrast.

In terms of visualized results, Fig. 4 shows the predicted grasps for Scene No.130 of the GraspNet-1Billion dataset. From the perspective of color depth and physical intuitive, the improved observation of the predicted grasps validates the robustness and effectiveness of our method.

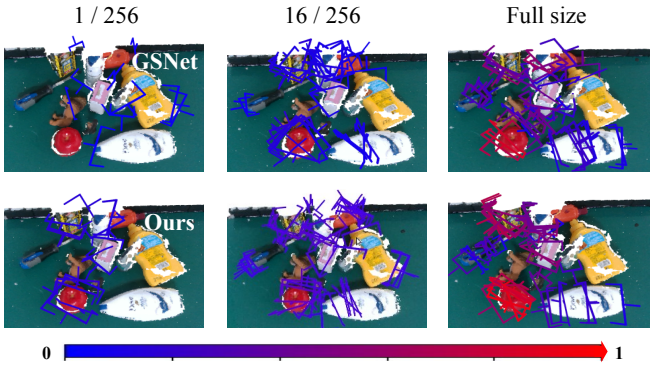

Fig. 4: Visualization of 6-DoF grasps predicted by GSNet (first row) and our method (second row) under varying training data sizes. Grasps are color-coded according to the predicted quality score, with higher values represented by deeper shades of red.

D. Ablation Studies

1) Guided Sampling Strategy: Ablation studies are first conducted on our guided false negative elimination strategy. Table II shows notable improvements resulting from the addition of the FNE strategy compared to the vanilla version of GSNet. By comparing the performance of the model with and without the FNE strategy, we gain insights into the improvements of 0.97/0.52 AP seen, 0.56/1.05 AP similar, 0.70/0.39 AP novel with 1/256 and 16/256 training data during the contrastive learning process. We attribute the results into FNE helps the network generalize better to unseen data without considering the mistakenly classified negative examples. These findings lay a solid foundation for the effectiveness of false negative elimination strategies in enhancing our contrastive learning framework.

TABLE II: Ablation studies of the FNE strategy.

Setting	Seen		Similar		Novel	
	1 / 256	16 / 256	1 / 256	16 / 256	1 / 256	16 / 256
GSNet	3.96	37.72	3.13	34.69	1.80	15.78
Ours w/o FNE	6.87	38.99	6.55	36.79	3.05	17.28
Ours (Full)	7.84	39.51	7.11	37.84	3.75	17.67

2) Data Augmentation: Data augmentation plays a crucial role in representation learning by enhancing the complexity of the training dataset. Table III explores the contributions of 3D data augmentations regarding to 1/256, 16/256 training frames in the pre-training phase. As evidenced by an average improvement of approximately 0.7 AP (Average Precision), the findings suggest that data augmentations are effective in enhancing downstream performance. Exposing models to diverse augmented data equips them to generalize to unseen data. This ultimately enhances their overall robustness and applicability across the grasp detection task.

TABLE III: Ablation studies of data augmentation.

Setting	AP Seen		AP Similar		AP Novel	
	1 / 256	16 / 256	1 / 256	16 / 256	1 / 256	16 / 256
GSNet	3.96	37.72	3.13	34.69	1.80	15.78
Ours w/o aug	7.19	38.76	6.74	36.48	3.43	17.19
Ours (Full)	7.84	39.51	7.11	37.84	3.75	17.67

E. Real Robotic Experiments

Physical Setup. The UR10e robotic arm is equipped with an eye-in-hand Realsense D435 camera to collect 1280×720 RGB-D images. In order to challenge the generalization ability of the model, we select all novel objects that do not appear in the training dataset with various shapes and sizes. We set up table-top scenes with three difficulty levels (e.g. easy, medium and hard) according to the density of object arrangement. Fig. 5 illustrates our experimental platform including the robotic arm, parallel-jaw gripper, camera, objects and plastic bin.

Grasp Execution. Specifically, we use ROS2 (Robot Operating System) to support communication of our robotic system. The operating system on the computer server is Ubuntu 20.04 LTS. The open-source software framework MoveIt is used for motion planning, collision checking, and control of robotic systems. After the hardware setup is complete, we adeptly execute the task procedure of picking objects from the clutter and placing them in the target bin.

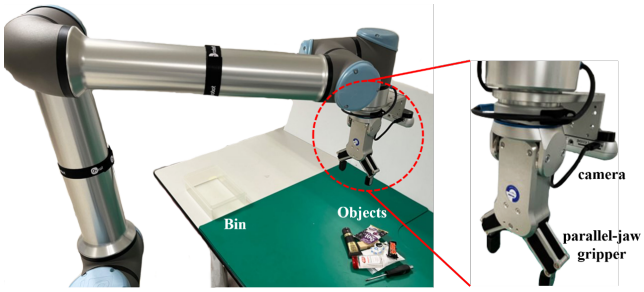

Fig. 5: Visualization of our experimental platform. The parallel-jaw gripper picks up objects one by one from the cluttered scene and places them into the dropping bin situated at a fixed position on the left side of the platform.

In the execution process, we consider a grasp attempt successful if the robot stably picks the object and places it in the plastic bin at the pre-defined location. The precision metrics success rate (SR) and completion rate (CR) are quantified as the percentage of successful grasping attempts and scene object completion. If one object fails to be picked up within three times, it will be manually removed and regarded as incomplete status. From the quantitative perspective, Table IV illustrates the grasping results of GSNet [14] and the version pre-trained by our GraspContrast in diverse cluttered scenes. We compare the success rates and object completion rates over five test scenes for easy, medium and hard level. It can be observed that GraspContrast helps achieve a better performance (SR: 91.47 vs. 85.93; CR: 99.17 vs. 96.67) in various scenarios compared to the vanilla GSNet without pre-training the backbone.

TABLE IV: Real experimental results

Level	Success Rate (%)		Completion Rate (%)	
	GSNet [14]	Ours	GSNet [14]	Ours
1	40/41	40/40	40/40	40/40
2	39/44	40/42	39/40	40/40
3	37/50	38/47	37/40	38/40
Total	116/135	118/129	116/120	118/120
Ratio	85.93%	91.47%	96.67%	98.33%

F. Conclusion

In this work, we propose GraspContrast - an easy-to-implement self-supervised framework that operates on unlabeled 3D point clouds to alleviate the need for manual annotation in the 6-DoF grasp detection task. Our framework leverages contrastive learning to learn effective representation by point-wise feature alignment. In the backbone pre-training process, we introduce the false negative elimination strategy to enhance feature diversity and mitigate the impact of false negative sample pairs. Experimental results on GraspNet-1Billion demonstrate that the fine-tuned model outperforms the version trained from scratch across various proportions of labeled frames. Our GraspContrast is applicable to various pipelines requiring 3D feature extraction. We anticipate that GraspContrast could inspire future work to consider the utilization of unlabeled data when developing intelligent algorithms in the field of robotic vision.

REFERENCES

[1] Edward Johns, Stefan Leutenegger, and Andrew 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.

[2] Jeffrey Mahler et al. "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics". In: *arXiv preprint arXiv:1703.09312* (2017).

[3] Martin Sundermeyer et al. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 13438–13444.

[4] Shaochen Wang, Zhangli Zhou, and Zhen Kan. "When transformer meets robotic grasping: Exploits context for efficient grasp detection". In: *IEEE robotics and automation letters* 7.3 (2022), pp. 8170–8177.

[5] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. "Perceiver-actor: A multi-task transformer for robotic manipulation". In: *Conference on Robot Learning*. PMLR. 2023, pp. 785–799.

[6] Julen Urain et al. "Se (3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 5923–5930.

[7] Hao-Shu Fang et al. "Graspnet-1billion: A large-scale benchmark for general object grasping". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11444–11453.

[8] Maximilian Gilles et al. "Metagraspnet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis". In: *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE. 2022, pp. 220–227.

[9] Ashish Jaiswal et al. "A survey on contrastive self-supervised learning". In: *Technologies* 9.1 (2020), p. 2.

[10] Longlong Jing and Yingli Tian. "Self-supervised visual feature learning with deep neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058.

[11] Xinlong Wang et al. "Dense contrastive learning for self-supervised visual pre-training". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 3024–3033.

[12] Ishan Misra and Laurens van der Maaten. "Self-supervised learning of pretext-invariant representations". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 6707–6717.

[13] Saining Xie et al. "Pointcontrast: Unsupervised pre-training for 3d point cloud understanding". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer. 2020, pp. 574–591.

[14] Chenxi Wang et al. "Graspness discovery in clutters for fast and accurate grasp detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 15964–15973.

[15] Andreas Ten Pas et al. "Grasp pose detection in point clouds". In: *The International Journal of Robotics Research* 36.13-14 (2017), pp. 1455–1473.

[16] Hongzhuo Liang et al. "Pointnetgpd: Detecting grasp configurations from point sets". In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, pp. 3629–3635.

[17] Charles R Qi et al. "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 652–660.

[18] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. "6-dof graspnet: Variational grasp generation for object manipulation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 2901–2910.

[19] Minghao Gou et al. "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 13459–13466.

[20] Zhixuan Liu et al. "TransGrasp: A Multi-Scale Hierarchical Point Transformer for 7-DoF Grasp Detection". In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 1533–1539.

[21] Aoran Xiao et al. "Unsupervised point cloud representation learning with deep neural networks: A survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).

[22] Richard Zhang, Phillip Isola, and Alexei A Efros. "Colorful image colorization". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 649–666.

[23] Mathilde Caron et al. "Deep clustering for unsupervised learning of visual features". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 132–149.

[24] Bruno Korbar, Du Tran, and Lorenzo Torresani. "Cooperative learning of audio and video models from self-supervised synchronization". In: *Advances in Neural Information Processing Systems* 31 (2018).

[25] Alon Faktor and Michal Irani. "Video segmentation by non-local consensus voting." In: *BMVC*. Vol. 2. 7. 2014, p. 8.

[26] Mang Ye et al. "Unsupervised embedding learning via invariant and spreading instance feature". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 6210–6219.

[27] Kaiming He et al. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[28] Ting Chen et al. "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[29] Ji Hou et al. "Exploring data-efficient 3d scene understanding with contrastive scene contexts". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 15587–15597.

[30] Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4d spatio-temporal convnets: Minkowski convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3075–3084.

[31] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[32] Alejandro Newell and Jia Deng. "How useful is self-supervised pretraining for visual tasks?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 7345–7354.

[33] Yueh-Cheng Liu et al. "Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining". In: *arXiv preprint arXiv:2104.04687* (2021).