# LiteGrasp: A Light Robotic Grasp Detection via Semi-Supervised Knowledge Distillation

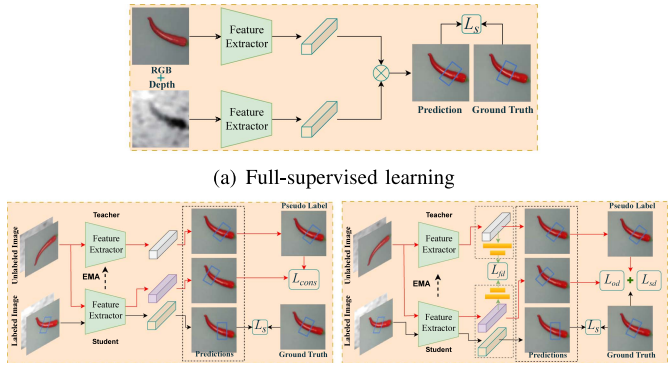
Linpeng Peng, Rongyao Cai, *Graduate Student Member, IEEE*, Jingyang Xiang, Junyu Zhu, Weiwei Liu, Wang Gao, and Yong Liu

***Abstract*—Grasping detection from single images in robotic applications poses a significant challenge. While contemporary deep learning techniques excel, their success often hinges on large annotated datasets and intricate network architectures. In this letter, we present LiteGrasp, a novel semi-supervised lightweight framework purpose-built for grasp detection, eliminating the necessity for exhaustive supervision and intricate networks. Our approach uses a limited amount of labeled data via a knowledge distillation method, introducing HRGrasp-Net, a model with high efficiency for extracting features and largely based on HRNet. We incorporate pseudo-label filtering within a mutual learning model set within a teacher-student paradigm. This enhances the transference of data from images with labels to those without. Additionally, we introduce the streamlined Lite HRGrasp-Net, acting as the student network which gains further distillation knowledge using a multi-level fusion cascade originating from HRGrasp-Net. Impressively, LiteGrasp thrives with just a fraction (4.3%) of HRGrasp-Net's original model size, and with limited labeled data relative to total data (25% ratio) across all benchmarks, regularly outperforming solely supervised and semi-supervised models. Taking just 6 ms for execution, LiteGrasp showcases exceptional accuracy (99.99% and 97.21% on Cornell and Jacquard data sets respectively), as well as an impressive 95.3% rate of success in grasping when deployed using a 6DoF UR5e robotic arm. These highlights underscore the effectiveness and efficiency of LiteGrasp for grasp detection, even under resource-limited conditions.**

***Index Terms*—Grasp detection, knowledge distillation, robotic grasping, semi-supervised learning.**

## I. Introduction

IN LIGHT of its substantial advantages in terms of automation and intelligence, robotic grasping has gained widespread application across industrial manufacturing and everyday life scenarios [1]. In the realm of sequential robotic grasping, the detection of a grasp is a pivotal aspect, enabling precise manipulation within real-world environments. However, achieving efficient and precise object grasp detection remains a challenging endeavor laden with practical complexities.




(a) Full-supervised learning

(b) Semi-supervised learning (c) Semi-supervised learning

Fig. 1. Comparison among different grasp detection methods.

The early techniques used primarily for grasp detection were heavily focused on recognized objects. For instance, Lenz et al. [2] introduced a novel representation for grasping that encompasses five dimensions. In contrast, Redmon et al. [3] significantly improved performance by optimizing multiple grasp candidates. Addressing limitations in generalization, Kumra et al. [4] introduced a multi-modal grasp predictor incorporating both RGB and depth channels. Nevertheless, these methods only achieved moderate success rates. However, these approaches exhibited constraints in scenarios involving unknown objects.

Recent efforts to enhance adaptability to novel scenarios including unknown objects have focused on data augmentation, innovative model architectures, and synthetic datasets [5], [6], [7], [8]. Noteworthy contributions include the SKGNet [6], which utilizes Selective Kernel convolution, and a transformer-based architecture [8] incorporating skip-connections. While these methods improved accuracy and generalization, their dependence on fully supervised training requiring large-scale labeled data limited real-world applicability (Fig. 1(a)).

To address the need for extensive labeled data, semi-supervised approaches [9], [10] utilized minimal labeled data and predominantly unlabeled data (Fig. 1(b)). [9] harnessed the power of mean-teacher semi-supervised learning, demonstrating competitive accuracy on the Cornell dataset [11]. Bai et al. [10] introduced an active semi-supervised strategy, effectively addressing insufficient real grasp labels. However, practical challenges arose due to large model sizes.

Manuscript received 3 April 2024; accepted 21 July 2024. Date of publication 31 July 2024; date of current version 7 August 2024. This article was recommended for publication by Associate Editor D. Seita and Editor A. Valada upon evaluation of the reviewers' comments. *(Corresponding authors: Wang Gao; Yong Liu.)*

Linpeng Peng, Rongyao Cai, Jingyang Xiang, Junyu Zhu, Weiwei Liu, and Yong Liu are with the College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: penglinpeng@zju.edu.cn; rycai@zju.edu.cn; jingyangxiang@zju.edu.cn; Junyu_Zhu@zju.edu.cn; 11932061@zju.edu.cn; yongliu@iipc.zju.edu.cn).

Wang Gao is with the Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing 100190, China (e-mail: gaowang@iipc.zju.edu.cn).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2024.3436336, provided by the authors.

Digital Object Identifier 10.1109/LRA.2024.3436336

To obtain a lightweight model without compromising performance, knowledge distillation techniques [12], [13] were employed to compress teacher models. Guo et al. [12] successfully distilled a deep teacher network for 6D pose estimation, achieving state-of-the-art results. Guan et al. [13] introduced a high-resolution 6D pose estimation network with knowledge distillation, showcasing superiority on the LINEMOD dataset.

To tackle the aforementioned challenges of insufficient real grasp annotations and redundancy network model parameters, motivated by the dual objectives of minimizing labeled data of above methods and ensuring a compact and efficient model for industrial applications, we propose a novel semi-supervised knowledge distillation method for grasp detection (Fig. 1(c)). Our approach involves employing HRNet [14] as the teacher feature extractor backbone and Lite-HRNet [15] as the student feature extractor backbone. A multi-level fusion cascade module facilitates information utilization between teacher and student networks. Comprehensive optimization is achieved through grasp detection, semi-supervised, and knowledge distillation losses. Experimental validation on three benchmark datasets, the Cornell Grasp Dataset [11], Jacquard Grasping Dataset [5] and Multi-object dataset [16], as well as in real-world grasp detection applications, demonstrates the effectiveness and efficacy of our approach. In summary, the primary contributions of this work are as follows:

1) We pioneer a novel approach for improving grasp detection by integrating semi-supervised and knowledge distillation. Our method efficiently utilizes both unlabeled and labeled data, streamlining the network through knowledge distillation. By this way, data labeling can be greatly reduced in real-world scenarios.
2) We introduce an innovative grasp detection approach using an efficient, lightweight HR-Net-based teacher-student network. It matches state-of-the-art performance while reducing model parameters by around 4.31% on two widely-recognized public datasets.
3) We develop an innovative multi-level fusion cascade module that spans from the teacher to the student network during feature knowledge distillation. This module effectively merges and bridges feature representation gaps between the two networks.
4) Comprehensive experiments confirm that our proposed framework delivers state-of-the-art performance across various domains, including public datasets such as Cornell, Jacquard and Multi-object dataset, as well as real-world robotic grasping scenarios.

## II. Related Work

### A. Grasp Detection

Grasp detection plays a pivotal role in robotic manipulation, involving the analysis of camera-captured images to identify optimal grasping positions. Previous approaches primarily rely on direct regression methods, often exploring feature extractor backbones based on regions of interest (ROI) [16], [17]. However, such methods may overlook intricate local features, leading to inadequate fusion. Recently, researchers have delved into strategies for enhancing network adaptability [5], [8], [18]. Presently, there is a burgeoning interest in methods striving to strike a balance between accuracy and inference speed through intricate yet lightweight architectures. For instance, EGNet [19] leverages feature maps from the bidirectional feature pyramid network (BiFPN) as input and generates grasp positions along with quality scores, demonstrating reduced parameter counts and enhanced detection accuracy. Furthermore, Cao et al. [20] introduce an efficient and lightweight generative structure for grasp detection networks, utilizing n-channel images as inputs and employing a Gaussian kernel-based grasp representation to encode training samples. This approach achieves outstanding equilibrium between accuracy and runtime performance. Despite advancing grasp detection, reliance on fully supervised training necessitates sufficient labeled data, potentially limiting real-world performance in data-scarce scenarios.

### B. Semi-Supervised Learning

Semi-supervised learning (SSL) is a powerful strategy for training models with limited labeled data, effectively incorporating unlabeled data. It finds applications across diverse domains, including image classification [21], [22], [23], medical image segmentation [24], 3D object detection [25], 6D object pose estimation [26], and grasp detection [9], [10]. MixMatch [22] introduced an algorithm predicting low-entropy labels for data-augmented unlabeled instances, integrating them with labeled data using MixUp. In response to complexity, Sohn et al. [23] introduced FixMatch, simplifying pseudo-label generation based on model predictions for weakly-augmented unlabeled images. By considering these, we integrate the semi-supervised learning to multi-level grasp detection, aiming at improving the training efficiency.

### C. Knowledge Distillation

In the domain of deep learning, the utilization of knowledge distillation emerges as a highly efficient technique extensively applied for constructive information transfer between networks during the training phase [27]. This method has diverse applications across various contexts [12], [13], [28]. For instance, task-oriented feature distillation was introduced by Zhang et al. [28], employing convolutional layers trained in a data-driven manner based on task loss, leading to significant enhancements in both image classification and 3D classification tasks. Zhu et al. [12] effectively solved the problem of 6D pose estimation assisted by distilling. Our LiteGrasp framework utilizes the advantages of both semi-supervised learning and knowledge distillation, resulting in significantly enhanced efficiency and effectiveness of the network model.

## III. Method

In this section, we will delve into a detailed explanation of our methodology. We begin by defining our problem, clearly elaborating the representation of robotic grasp detection in Subsection A. Moving onwards to Subsection B, we introduce the HRGrasp-Net feature extractor, otherwise known as HRNet [14]. Subsequently, in Subsection C, we present a semi-supervised training framework for grasp detection that leverages data augmentation

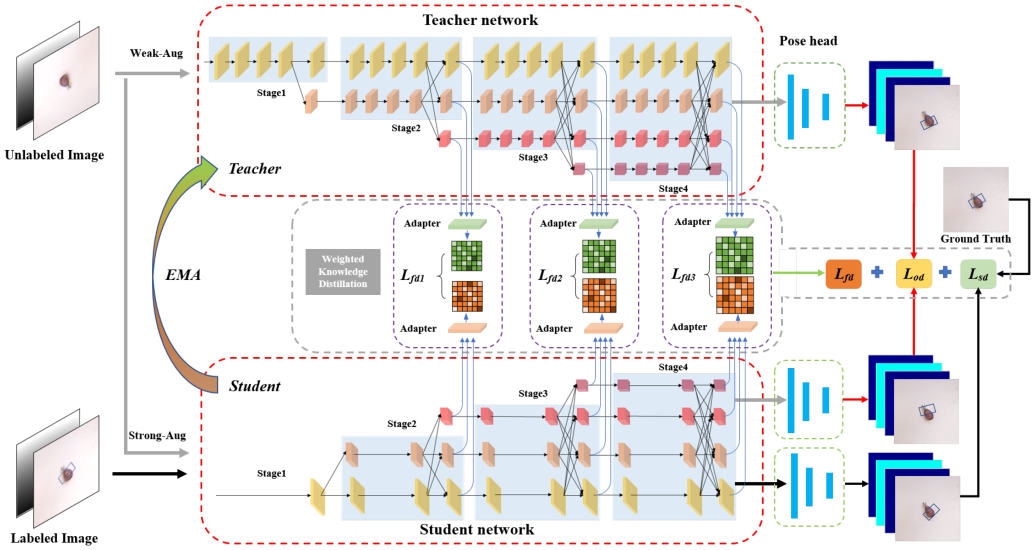



Fig. 2. Overview of the framework. Our model takes as input the image captured by the camera and generates a pixel-level grasp representation.

within a teacher-student model. Concluding this section, Subsection D elucidates the application of knowledge distillation in a cascade-like manner with weighted steps to improve grasp detection performance.

### A. Robotic Grasp System Representation

The model processes RGB and depth images as inputs to generate grasp candidates, as initially defined in the grasp detection problem presented in [2], [3]. Following the approach of [6], [7], [29], [30], we employ an enhanced grasp representation, deviating from the original five-dimensional grasp rectangle [2], [3], to articulate the grasp pose within the 2D image space:

$$g_i = \{\boldsymbol{p}, \theta, w, q\} \tag{1}$$

where $\boldsymbol{p}$ represents the position of the center point expressed in image coordinate frame as $\boldsymbol{p} = (u, v)$. The variables $\theta$ and $w$ denote the grasp angle around the z-axis and the grasp width of the gripper, respectively, while $q$ indicates the grasp quality.

Following eye-in-hand calibration, the camera coordinates are transformed into the robot frame, and additional transformation from image coordinates to camera coordinates is accomplished through camera calibration. Similiar to [7], [29] the grasp map in robot space is then defined as:

$$G_r = {}^{R}T_{C}\,{}^{C}T_{I}(g_i). \tag{2}$$

where ${}^{R}T_{C}$ represents the transformation matrix from camera coordinate frames to robot coordinate frames, and ${}^{C}T_{I}$ signifies the transformation matrix from 2D image coordinate space to camera coordinate frames.

### B. HR-Net Grasp Detection Network Architecture

We present HRGrasp-Net, a two-step network designed for object grasp detection based on HR-Net, as depicted in Fig. 2 (Teacher network as example). HRGrasp-Net comprises two main components: multi-scale fusion feature extraction

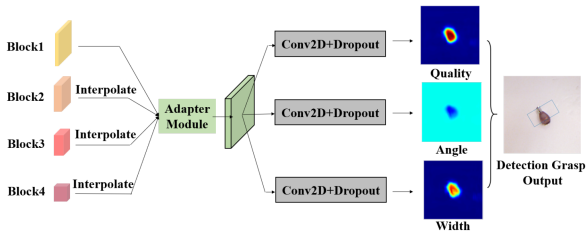



Fig. 3. The architecture of grasp detection pose head module.

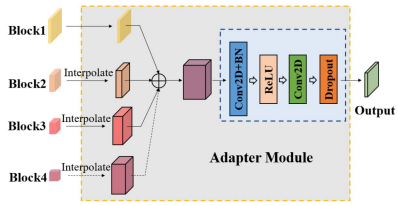



Fig. 4. The architecture of adapter module. (The dotted-line block4 branch is specifically designed for the fusion of stages 3 and 4.).

by the HRNet backbone and acquisition of object grasp pose through the pose head module. The HRNet backbone utilizes four stages for concurrent high-to-low resolution feature fusion, with each stage containing multi-resolution features generated by others. During the feature extraction process, HRGrasp-Net takes 224 × 224 (320 × 320) RGB-D images as inputs and produces four high-to-low resolution fusion feature maps through the operations of the last stages in HRNet.

Following the feature extraction by HRNet, a pose head module is introduced to handle the four high-to-low resolution fusion feature maps. The architecture of the pose head module is illustrated in Fig. 3, incorporating an adapter module (depicted in Fig. 4 and will be explained in Subsection D) for feature map alignment and fusion. Additionally, three Conv2D and Dropout layers are employed to generate grasp detection predictions for

three branches, encompassing grasp quality, angle, and width predictions.

For HRGrasp-Net training, we devise a supervised training loss aimed at minimizing the disparity between prediction results and ground truth labels. Specifically, we train HRGrasp-Net to learn the mapping function $F : I \rightarrow G$, where $I = \{I_1, \ldots, I_n\}$ represents input images, and $G = \{g_1^p, \ldots, g_n^p\}$ denotes the generated predictions. The supervised loss is designed to constrain $G$ with corresponding ground truth labels $G_{gt} = \{g_1^{gt}, \ldots, g_n^{gt}\}$. In this context, the single-item supervised loss $L_{s(j)}$ is defined as(j can be $\theta$, w and q):

$$L_{s(j)} = \frac{1}{N} \sum_{i=1}^{N} Smooth_{L1}(g_i^p - g_i^{gt}) \tag{3}$$

where $N$ represents the count of grasp candidates and the $Smooth_{L1}$ loss is formulated as:

$$Smooth_{L1} = \begin{cases} 0.5(g_i^p - g_i^{gt})^2, & \text{if } |g_i^p - g_i^{gt}| < 1, \\ |g_i^p - g_i^{gt}| - 0.5, & \text{otherwise} \end{cases} \tag{4}$$

Given that pose head module outputs three prediction branches, namely grasp quality, angle, and width prediction, the loss function $L_{s(j)}$ is employed to optimize each prediction branch. Furthermore, angle prediction is computed in the form of $(\cos(2\theta), \sin(2\theta))$, as detailed in [6], [7], [29], [30]. Consequently, the total supervised loss $L_s$ is expressed as:

$$L_s = \delta_1 L_{s(\theta)} + \delta_2 L_{s(w)} + \delta_3 L_{s(q)}. \tag{5}$$

where $\delta_1$, $\delta_2$ and $\delta_3$ are followed in [6] for safe grasping, which we set to 1.0, 1.0 and 1.2, respectively.

### C. Data Augmentation Based Semi-Supervised Learning

The introduced semi-supervised process, depicted in Fig. 2, is inspired from the MixMatch framework [22]. MixMatch is a comprehensive semi-supervised approach characterized by data augmentation on mixed labeled and unlabeled data using MixUp. In this context, we integrate a teacher-student mutual learning method into the MixMatch framework. For our study, we employ HRNet-W18 and HRNet-W48 [14] (two different model parameters just for comparison) as backbone feature extraction networks.

The solution we propose comprises of two training phases. Initially, there is a pre-training phase where we train the teacher network using supervised learning on labeled images. Following that, we proceed to the semi-supervised learning (SSL) phase, wherein both the teacher and student networks are trained using unsupervised learning on data-augmented unlabeled images.

To be specific, we initiate the training by supervising the teacher network to get teacher model using input images with ground truth labeled data, denoted as $D^l = \{(x_1^l, g_1^l), \ldots, (x_n^l, g_n^l)\}$. Subsequently, for one batch of labeled input image data and another batch of unlabeled input image data, we employ weak and strong data augmentation techniques to get mixed data, followed by mixup, similar with the framework of MixMatch, represented as $D^m = \{(x_1^m, g_1^m), \ldots, (x_n^m, g_n^m)\}$. Additionally, in the step of strong data augmentation, we use teacher network to generate pseudo-labels as ground truth labels for mixup. The teacher and student networks share identical configurations, and an exponential moving average (EMA) method is employed to update teacher weights based on the student model.

The ultimate semi-supervised loss is formulated as:

$$L_{ssp} = L_l + \gamma L_u. \tag{6}$$

where $L_l$ and $L_u$ are the supervised loss and the unsupervised loss, calculated by formula (4); $\gamma$ denotes the unsupervised loss weight scale and to be set 1.0 in this letter.

### D. Weighted Multi-Step Knowledge Distillation

As depicted in Fig. 2, our knowledge distillation framework comprises a highly-performing teacher network $T$ and a lightweight student network $S$. In this study, we designate HRNet-W18 and HRNet-W48 as the teacher feature extraction networks, while Lite-HRNet-18-W18 and Lite-HRNet-18-W48 serve as the student feature extraction networks. The teacher and student feature extraction networks are followed by a pose head module, collectively forming the networks $T$ and $S$. Specifically, Lite-HRNet [15] is a lightweight model designed to optimize the parameters of HRNet, facilitating its practical application in industrial scenarios. Lite-HRNet-18-W18 and Lite-HRNet-18-W48 share the same architecture based on Lite-HRNet but have differences in network layers. Similar architectural sharing exists between HRNet-W18 and HRNet-W48.

Due to its computational efficiency, stability, and ease of implementation, we adopt offline knowledge distillation in this context [27]. Specifically, we utilize the upper stage SSL-trained model as the teacher model and employ the more lightweight Lite-HRNet-18-W18 and Lite-HRNet-18-W48, each equipped with a pose head module, as the student models. In this study, we establish two sets of distillation schemes: Lite-HRNet-18-W18 is aligned with HRNet-W18 due to their matching output channels, and similarly, Lite-HRNet-18-W48 is aligned with HRNet-W48. Additionally, we denote the distilled models obtained from Lite-HRNet-18-W18 and Lite-HRNet-18-W48 as Lite-GraspNet-W18 and Lite-GraspNet-W48.

Taking inspiration from [28], we employ feature distillation and logit distillation to refine the student model based on the teacher model. Using Lite-HRNet-18-W18 and HRNet-W18 as an example, in the feature distillation step, we propose a weighted multi-step knowledge distillation approach to address unbalanced feature transfer across different stages during the feature extraction process in both the teacher and student networks. Specifically, we first use the adapter module (showed in Fig. 4) to align the output feature maps in stages 2, 3, and 4 for HRNet-W18 and Lite-HRNet-18-W18. Subsequently, different training weight scales are set to adjust the contribution for precision performance across the three stages.

In another step of logit distillation, we utilize data-augmented images from the SSL process to train both the teacher and student networks, obtaining teacher and student output prediction maps. By comparing these two output maps, we can identify the desired differences.

Let $L_{fdi}$ $(i = 1, 2, 3)$ be the loss of feature distillation, $L_{od}$ be the loss of logit distillation, and $L_{sd}$ be the loss of supervised loss. $L_{od}$ and $L_{sd}$ are calculated by Smooth L1 loss (same as

formula (3)) while $L_{fdi}$ by KL divergence. The total weighted multi-step loss of distillation is defined as:

$$L_{dist} = L_{sd} + \alpha L_{od} + \beta L_{fd} \tag{7}$$

where $\alpha$ and $\beta$ are balanced weight scales, which we set to 5.0 and 1.0, respectively.

The weighted feature distillation loss can be defined as:

$$L_{fd} = \lambda_1 L_{fd1} + \lambda_2 L_{fd2} + \lambda_3 L_{fd3}. \tag{8}$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are feature balanced weight scales, which we set to 0.1, 0.3 and 0.6, respectively.

## IV. Experiments

In this section, we perform comprehensive comparative experiments to assess the performance of Lite-Grasp. Our evaluation includes testing our proposed method on different publicly available grasp datasets: Cornell [11], Jacquard [5] and a multi-object dataset [16]. We also conduct ablation studies to investigate the influence of various network components. Furthermore, we validate the effectiveness of Lite-Grasp using both simulation and a real UR5e robotic manipulator. The code is released at https://github.com/ZJU-PLP/LiteGrasp.

### A. Datasets and Experimental Setup

*1) Datasets and Evaluation Metrics:* Consistent with previous studies [7], [18], [29], we employ the same evaluation metrics to assess our model's performance on the grasp datasets, including Cornell [11], Jacquard [5], and the multi-object dataset [16]. In particular, a predicted grasp is considered correct when it satisfies the following two conditions:

- Angle difference: The difference between the predicted grasping angle and ground truth should not greater than 30°.
- Jaccard index: The Jaccard index between the predicted grasp $g_p$ and the corresponding grasp label $g_{gt}$ should exceed 25%, which is formulated in the following:

$$J(g_p, g_{gt}) = \frac{|g_p \cap g_{gt}|}{|g_p \cup g_{gt}|}. \tag{9}$$

*2) Experiment Setup:* The entire system is implemented on an Ubuntu 18.04 computer system equipped with a singular NVIDIA TITAN RTX possessing 24 GB of memory. Implementation utilizes the PyTorch 1.13 deep learning framework in conjunction with CUDA 11.6 packages. Employing the Adam optimizer with a learning rate of 0.001 and a batch size set at 8, the network undergoes distinct training stages, encompassing full-supervised learning, semi-supervised learning, and knowledge distillation training. Followed the dataset distribution of [6] and [29], 90% is allocated to training, and the remaining 10% is designated for testing. Within the training subset, 90% is utilized for actual training, while the remaining 10% is dedicated to validation.

### B. Experiments and Analysis on Cornell Dataset

The Cornell grasping dataset comprises 885 images, each with a resolution of 640 × 480. All images in the dataset are resized to 224 × 224 and subjected to various data augmentation techniques to prevent overfitting when fed into the network. To ensure fairness, the evaluation follows both image-wise (IW) and object-wise (OW) settings, consistent with previous studies [6], [7], [29].

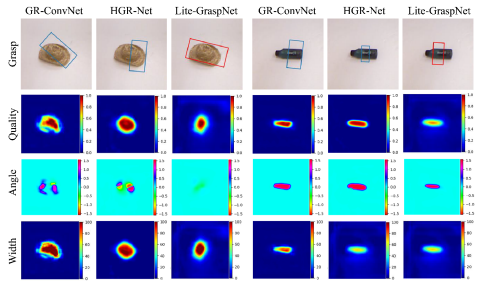



Fig. 5. Comparison studies on Cornell dataset.

TABLE I
The Accuracy on Cornell Grasping Dataset

| Category | Author | Method | Year | Input modality | Input size | Accuracy (%) | |
|---|---|---|---|---|---|---|---|
| | | | | | | IW | OW |
| FSL | Redmon [3] | MultiGrasp | 2015 | RGB-D | 224x224 | 88.0 | 87.1 |
| | Kumra [4] | ResNet-50 | 2017 | RGB-D | 224x224 | 89.2 | 88.9 |
| | Asif [17] | GraspNet | 2018 | RGB-D | 224x224 | 90.2 | 90.6 |
| | Chu [16] | FasterR-CNN | 2018 | RG-D | 227x227 | 96.0 | 96.1 |
| | Morrison [31] | GG-CNN2 | 2019 | D | 300x300 | 84.0 | 82.0 |
| | Kumra [29] | GR-ConvNet | 2020 | RGB-D | 300x300 | 97.7 | 96.6 |
| | Wu [32] | AFGD | 2021 | RGB-D | 320x320 | 99.4 | 98.9 |
| | Yu [18] | SE-ResUNet | 2022 | RGB-D | – | 98.2 | 97.1 |
| | Kumra [33] | GR-ConvNet v2 | 2022 | RGB-D | 300x300 | 98.8 | 97.7 |
| | Xu [34] | GKNet | 2022 | RGB-D | – | 96.9 | 95.7 |
| | Wang [8] | TF-Grasp | 2022 | RGB-D | 224x224 | 97.99 | 96.7 |
| | Yu [6] | SKGNet | 2022 | RGB-D | 288x288 | 99.1 | 98.4 |
| | Cao [20] | Efficient Grasp | 2022 | RGB-D | 224x224 | 97.8 | – |
| | Wu [35] | PLG-Net | 2023 | RGB-D | 300x300 | – | 98.1 |
| | Zhou [7] | AAGDN | 2023 | RGB-D | 320x320 | 99.3 | 98.8 |
| | Zhou [30] | HRG-Net | 2023 | RGB-D | 224x224 | 99.5 | 97.5 |
| SSL | Zhu [9] | – | 2020 | RGB-D | 288x288 | 97.4 | – |
| | Shukla [36] | RGGCNN2 | 2023 | RGB | – | 95.5 | – |
| | Ours | Lite-GraspNet-W48(lb=25%) | – | RGB | 224x224 | 97.75 | 96.63 |
| | | | | D | | 98.87 | 97.75 |
| | | | | RGB-D | | 99.99 | 97.75 |

To emphasize the merits of our methodology, we perform comparative analyses involving established models such as GR-ConvNet [29] and the recent HRG-Net [30]. As depicted in Fig. 5, we showcase the results of grasp detection for objects not encountered during training. In our evaluation procedure, we designate the pixel with the highest quality score as the optimal grasp pose. The outcomes unequivocally reveal that Lite-Grasp surpasses both GR-ConvNet and HRG-Net, attaining the highest grasp quality score and illustrating the competitive prowess of our approach.

A more extensive examination of grasp detection accuracy is conducted, comparing our method with existing algorithms, and the results are detailed in Table I. Lite-GraspNet-W48 attains the highest accuracy in the IW split and exhibits competitive accuracy in the OW split, with values of 99.9% and 97.75%, respectively, underscoring its state-of-the-art performance. Moreover, our approach sustains competitive performance, even when trained with only 25% labeled data, affirming its practical applicability.

### C. Experiments and Analysis on Jacquard Dataset

The Jacquard grasp dataset comprises 54,000 RGB-D images and 1.1 million successful grasp rectangles, all generated within a simulated physical environment. Given the dataset's size, no data augmentation is applied in this case.

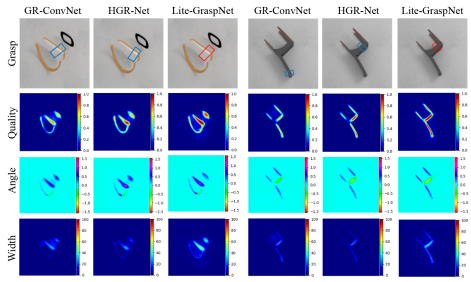



Fig. 6. Comparison studies on Jacquard dataset.

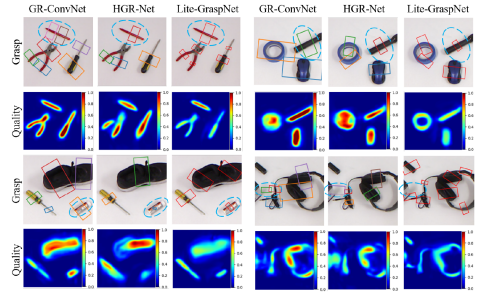



Fig. 7. Comparison studies on the multi-object dataset.

TABLE II
THE ACCURACY ON JACQUARD GRASPING DATASET

| Author | Method | Year | Input modality | Input size | Accuracy (%) |
|---|---|---|---|---|---|
| Depierre [5] | Jacquard | 2018 | RGB-D | 227x227 | 74.2 |
| Morrison [31] | GG-CNN2 | 2019 | D | 300x300 | 84.0 |
| Kumra [29] | GR-ConvNet | 2020 | RGB-D | 300x300 | 94.6 |
| Wu [32] | AFGD | 2021 | RGB-D | 320x320 | 96.2 |
| Yu [18] | SE-ResUNet | 2022 | RGB-D | – | 95.7 |
| Kumra [33] | GR-ConvNet v2 | 2022 | RGB-D | 300x300 | 95.1 |
| Xu [34] | GKNet | 2022 | RGB-D | – | 96.99 |
| Wang [8] | TF-Grasp | 2022 | RGB-D | 224x224 | 94.6 |
| Yu [6] | SKGNet | 2022 | RGB-D | 288x288 | 95.9 |
| Cao [20] | Efficient Grasp | 2022 | D | 227x227 | 95.6 |
| Wu [35] | PLG-Net | 2023 | RG-D | 300x300 | 96.0 |
| Zhou [7] | AAGDN | 2023 | RGB-D | 320x320 | 96.2 |
| Zhou [30] | HRG-Net | 2023 | RGB-D | 224x224 | 96.5 |
| Ours | Lite-GraspNet-W48(lb=25%) | – | RGB | 320x320 | 96.02 |
| | | | D | | 97.05 |
| | | | RGB-D | | 97.21 |

TABLE III
IMPACT OF DIFFERENT KNOWLEDGE DISTILLATION COMPONENTS

| NetWork Settings | Accuracy(%) |
|---|---|
| (a) HRNet-W18 | 93.71 |
| (b) HRNet-W18+SSL | 96.57 |
| (c) HRNet-W18+SSL+SD | 95.27 |
| (d) HRNet-W18+SSL+SD+FD | 96.64 |
| (e) HRNet-W18+SSL+SD+FD+OD | **96.84** |

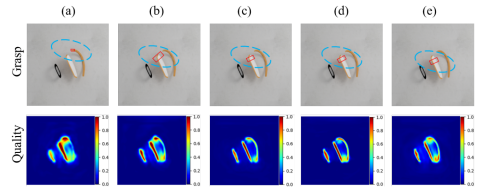



Fig. 8. Visual comparisons on different components.

We also perform comparative experiments on the Jacquard dataset for unseen objects, as depicted in Fig. 6. Our method consistently yields more accurate grasp rectangles, irrespective of the object's size, in comparison to other methods. This improved performance can be attributed to our model's ability to learn well-fused feature information across multiple scales of the object. These experiments underscore the applicability and robustness of our method for objects with diverse shapes.

Quantitative results regarding grasp detection accuracy are presented in Table II. Only with depth data as input, our approach achieves outstanding performance, boasting a detection accuracy of 97.05%. This result surpasses existing methods and gets the second-top position on the Jacquard dataset.

### D. Experiments and Analysis on Multi-Object Dataset

To evaluate Lite-GraspNet's efficacy in cluttered scenes, we conduct extensive experiments utilizing the Multi-object dataset. Comprising 97 RGB-D images, each scene incorporates a minimum of three previously unseen objects. Adhering to standard practices, the objects used for testing are novel to the network. Fig. 7 presents detailed comparative detection results, featuring GR-ConvNet and HRG-Net.

These results demonstrate that Lite-Grasp's multi-scale feature fusion improves the accuracy of robotic grasp detection by extracting information from image data. From above experimental results, we conclude that knowledge distillation is beneficial to small objects grasp detection in cluttered scenes. However, the benefits are relatively small to deal with large objects from knowledge distillation maybe the student model can already learn the key feature information.

### E. Ablation Studies

*Influence of the different components:* To assess the impact of various components in our distillation scheme for grasp pose learning, we perform an ablation study on the Jacquard dataset using depth data as input for Lite-GraspNet-W18(l b = 25%). As shown in Table III, ($\alpha = 0.0$ and $\beta = 0.0$ in (7)) denotes only using logit(output) distillation, and ($\alpha > 0.0$ and $\beta > 0.0$ in (7)) means using logit along with feature-similarity distillation. We also visualize the grasp detection results on different network settings in Fig. 8.

From Table III, we can conclude that the model benefits when both output-distillation and feature-similarity distillation are employed, resulting in a significant improvement of 1.57% in accuracy. Fig. 8 also portrays the trend that as the count of functional modules escalates, the detection results progressively enhance in terms of accuracy. These findings demonstrate that the weighted distillation schemes, with their different components, collectively contribute to enhancing grasp detection accuracy.

*Comparison of different data label amounts:* To delve deeper into the effectiveness of the Lite-GraspNet network, we assess how leveraging unlabeled images impacts the performance of our model with varying amounts of labeled samples. We conduct these evaluations on training datasets from both Cornell(OW split) and Jacquard using depth data as input, testing the proposed model's performance when using a range of labeled data percentages, from 12.5% to 100%.

TABLE IV
LITE-GRASPNET UTILIZES CORNELL AND JACQUARD DATASETS WITH VARYING LABELED DATA RATIOS FOR DEPTH INPUT

| Label Amounts | Learning Framework | Backbone | | Accurracy(%) | | | |
|---|---|---|---|---|---|---|---|
| | | | | Cornell | | Jacquard | |
| 12.5% | FSL | HRNet-W18 | Lite-HRNet-18-W18 | 93.26 | 93.26 | 96.40 | 93.96 |
| | | HRNet-W48 | Lite-HRNet-18-W48 | 95.51 | 94.38 | 95.91 | 94.16 |
| | SSL | HRNet-W18 | | 95.51(↑ 2.25) | | 96.53(↑ 0.13) | |
| | | HRNet-W48 | | 96.63(↑ 1.12) | | 96.22(↑ 0.31) | |
| | SSL+KD | HRNet-W18 + Lite-HRNet-18-W18 | | 95.51(↑↑ 0.00) | | 96.43(↓↓ 0.10) | |
| | | HRNet-W48 + Lite-HRNet-18-W48 | | 96.63(↑↑ 0.00) | | 96.02(↓↓ 0.20) | |
| 25% | FSL | HRNet-W18 | Lite-HRNet-18-W18 | 94.38 | 93.26 | 96.02 | 93.71 |
| | | HRNet-W48 | Lite-HRNet-18-W48 | 95.51 | 94.38 | 95.74 | 94.88 |
| | SSL | HRNet-W18 | | 95.51(↑ 0.00) | | 96.57(↑ 0.52) | |
| | | HRNet-W48 | | 97.75(↑ 2.24) | | 96.59(↑ 0.85) | |
| | SSL+KD | HRNet-W18 + Lite-HRNet-18-W18 | | 96.63(↑↑ 0.24) | | 96.84(↑↑ 0.27) | |
| | | HRNet-W48 + Lite-HRNet-18-W48 | | 97.75(↑↑ 0.00) | | 97.05(↑↑ 0.46) | |
| 50% | FSL | HRNet-W18 | Lite-HRNet-18-W18 | 98.88 | 97.75 | 97.56 | 96.15 |
| | | HRNet-W48 | Lite-HRNet-18-W48 | 99.99 | 98.88 | 97.45 | 96.04 |
| | SSL | HRNet-W18 | | 99.99(↑ 1.11) | | 97.59(↑ 0.03) | |
| | | HRNet-W48 | | 99.99(↑ 0.00) | | 97.83(↑ 0.38) | |
| | SSL+KD | HRNet-W18 + Lite-HRNet-18-W18 | | 99.99(↑↑ 0.00) | | 96.37(↑↑ 0.04) | |
| | | HRNet-W48 + Lite-HRNet-18-W48 | | 99.99(↑↑ 0.00) | | 97.35(↓↓ 0.24) | |
| 100% | FSL | HRNet-W18 | Lite-HRNet-18-W18 | 99.99 | 99.99 | 97.41 | 96.37 |
| | | HRNet-W48 | Lite-HRNet-18-W48 | 99.99 | 99.99 | 97.45 | 96.62 |
| | SSL | HRNet-W18 | | – | | – | |
| | | HRNet-W48 | | – | | – | |
| | SSL+KD | HRNet-W18 + Lite-HRNet-18-W18 | | 99.99(↑↑ 0.00) | | 96.86(↓↓ 0.55) | |
| | | HRNet-W48 + Lite-HRNet-18-W48 | | 99.99(↑↑ 0.00) | | 97.50(↑↑ 0.05) | |

TABLE V
EFFICIENCY COMPARISON OF DIFFERENT METHODS (APPROX)

| Methods | Params(M) | | Pruning Scale(%) | Speed(ms) | Accurracy(%) | Hardware (GPU) |
|---|---|---|---|---|---|---|
| Morrison | 0.071 | | \ | 200-500 | 84.0 | GeForce GTX 1070 |
| Kumura | 1.9 | | \ | 7 | 94.6 | GeForce GTX 1080 Ti |
| Cao | 1.2 | | \ | 6 | 95.6 | GeForce RTX 2080 Ti |
| Wang | 6.8 | | \ | 41.6 | 94.6 | GeForce RTX 3090 |
| Zhou | 65.86 | | \ | 53.7 | 96.5 | GeForce RTX 2080 Super |
| | Teacher backbone | Student backbone | \ | | | |
| Ours(Ib=25%) | ResNet-50(25.6) | ResNet-18(11.2) | 43.75 | 43.2(1324.5) | 93.32 → 94.31(↑ 0.99) | TITAN RTX(CPU) |
| | ResNet-101(44.5) | ResNet-18(11.2) | 25.17 | 43.2(1324.5) | 93.32 → 94.53(↑ 1.21) | |
| | ResNet-152(60.2) | ResNet-18(11.2) | 18.60 | 43.2(1324.5) | 93.32 → 94.71(↑ 1.39) | |
| | HRNet-W48(66.53) | Lite-HRNet-W48(2.71) | 4.07 | 6.44(348.2) | 94.88 → 97.05(↑ 2.17) | |
| | HRNet-W18(9.74) | Lite-HRNet-W18(0.42) | 4.31 | 1.88(68.9) | 93.71 → 96.84(↑ 3.13) | |
| | HRNet-W48(66.53) | Lite-HRNet-W18(0.42) | 0.62 | 1.88(68.9) | 93.71 → 96.73(↑ 3.02) | |

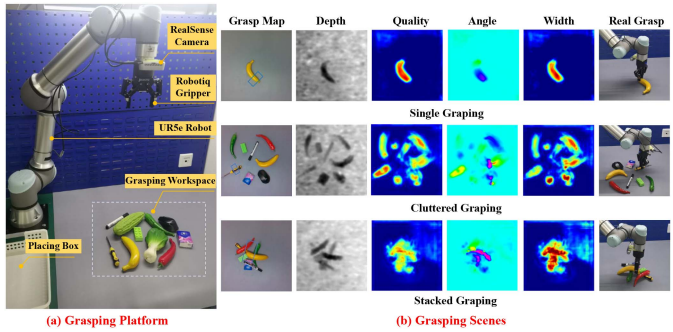



Fig. 9. Grasp experiments of the real UR robot.

TABLE VI
GRASP SUCCESS RATES IN REAL-WORLD SCENES

| Method | Physical grasp | Success rate(%) |
|---|---|---|
| GR-ConvNet [29] | 427/450 | 94.9 |
| Efficient Grasp [20] | 472/526 | 89.7 |
| TF-Grasp [8] | 152/165 | 92.1 |
| AAGDN [7] | 544/575 | 94.6 |
| Ours | 343/360 | 95.3 |

Table IV clearly illustrates that, for varying amounts of labeled samples, the semi-supervised learning framework (SSL) consistently outperforms the full-supervised learning framework (FSL) in terms of final detection accuracy, regardless of the backbone settings. This observation underscores the effectiveness of the proposed SSL method.

While there may be a slight sacrifice in accuracy after knowledge distillation in a few cases, the combination of semi-supervised learning and knowledge distillation (SSL+KD) achieves a competitive grasp detection result of 96.43% on the Jacquard dataset, even with just 12.5% labeled data. This result highlights the model's ability to effectively distill valuable information from the teacher network to the student network.

*Comparison of network efficience:* To confirm the efficiency and effectiveness of our proposed model, we analyzed our model's parameters, FLOPs(floating-point operations), and inference time, juxtaposing them with several prevalent methods. We interpreted the trend connected with accuracy performance through the calculation of the pruning scale, denoting the percentage ratio of the student model's participation relative to the teacher model's parameters. To ensure our framework's adaptability, we tested it in different frequently used feature extractor backbones.

Our teacher network GraspNet (GraspNet-W18 and GraspNet-W48) is similar with HRG-Net [30] while student network GraspNet(Lite-GraspNet-W18 and Lite-GraspNet-W48) is similar with Lite-HRNet [15]. Thanks to knowledge distillation, we've designed a lightweight grasp detection architecture that not only achieves superior detection accuracy but also runs faster in different backbones. Specifically, we achieve a reduction in the model parameters of the student network to approximately 4.3% compared to the teacher network while maintaining competitive performance levels. As shown in Table V, it's evident that our method Lite-GraspNet-W18 has only 0.42 million parameters and 0.71 billion FLOPs, which are significantly lower than other methods. Remarkably, even with a mere 0.6% participation of teacher model parameters, our method can still attain a comparative level of accuracy performance. Furthermore, the model's inference time is a mere 1.88 milliseconds under NVIDIA Titan RTX hardware settings(68.9 milliseconds under CPU setting). The experimental results clearly demonstrate that the proposed method delivers exceptional efficiency when executed on GPU hardware, making it suitable for industrial grasping applications.

### F. Experiments on Real-World Environment

In this section, we delve into the robotic grasping experiments and the outcomes obtained in real-world scenarios. To ensure a fair comparison, we employ an open-loop grasping method akin to prior research [6], [20], [29] and assess our approach in the following contexts: i) single objects, ii) cluttered objects, and iii) stacked objects.

To further assess the accuracy and generalization of Lite-GraspNet in real-world scenarios, we apply it to grasping tests with a real UR5e robot equipped with a Robotiq 2F-85 gripper. A RealSense D435 RGB-D camera is utilized as the perceptual component, mounted in an eye-in-hand configuration. In our robotic manipulation scene, a grasp is deemed successful if the robot can adeptly lift an object and accurately place it in the designated target box. As depicted in Fig. 9, the trained model is tested in three distinct real-world scenes, and the success rates are detailed in Table VI (More details can be found in

our attached video). The robot, equipped with Lite-GraspNet, has accomplished a total of 360 grasps in different scenes, achieving a remarkable grasp success rate of 95.3%. These results underline the effectiveness of the proposed method, which combine semi-supervised knowledge distillation to refine grasp performance and consistently achieve successful grasps in real robotic applications.

## V. Discussion and Conclusion

This letter introduces LiteGrasp, a novel framework designed for semi-supervised lightweight grasp detection. It addresses challenges about excessive supervision and complex model architectures by leveraging a limited set of labeled data and implementing knowledge distillation based on HR-Net. Emphasizing multi-scale features, LiteGrasp facilitates effective information transfer from the teacher to the student networks. Experimental evaluations on established datasets demonstrate competitive performance. Robustness is validated through real-world grasp experiments, featuring the UR robot arm and Robotiq parallel-jaw gripper. On the other hand, our methodology predominantly concentrates on achieving superior accuracy and efficiency in the realm of close-vocabulary grasp detection, which presents fewer complexities as opposed to open-vocabulary situations. Hedging the path for future explorations, we will delve into more universally applicable grabbing approaches, integrating Vision-Language Models to facilitate generic robotic grasping and combat high-challenge open-vocabulary grasp detection scenarios. Further precision and success rates will also be pursued by investigating closed-loop grasping strategies.

## References


[1] R. Newbury et al., "Deep learning approaches to grasp synthesis: A review," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3994–4015, Oct. 2023.

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

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

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

[5] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in *2018 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 3511–3516.

[6] S. Yu, D.-H. Zhai, and Y. Xia, "SKGNet: Robotic grasp detection with selective kernel convolution," *IEEE Trans. Automat. Sci. Eng.*, vol. 20, no. 4, pp. 2241–2252, Oct. 2023.

[7] Z. Zhou, X. Zhu, and Q. Cao, "AAGDN: Attention-augmented grasp detection network based on coordinate attention and effective feature fusion method," *IEEE Robot. Automat. Lett.*, vol. 8, no. 6, pp. 3462–3469, Jun. 2023.

[8] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 8170–8177, Jul. 2022.

[9] H. Zhu et al., "Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation," in *2020 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 9608–9613.

[10] F. Bai, D. Zhu, H. Cheng, P. Xu, and M. Q.-H. Meng, "Active semi-supervised grasp pose detection with geometric consistency," in *2021 IEEE Int. Conf. Robot. Biomimetics*, pp. 1402–1408.

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

[12] S. Guo, Y. Hu, J. M. Alvarez, and M. Salzmann, "Knowledge distillation for 6D pose estimation by aligning distributions of local predictions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 18633–18642.

[13] Q. Guan, Z. Sheng, and S. Xue, "HRPose: Real-time high-resolution 6D pose estimation network using knowledge distillation," *Chin. J. Electron.*, vol. 32, no. 1, pp. 189–198, 2023.

[14] J. Wang et al., "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[15] C. Yu et al., "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10440–10450.

[16] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.

[17] U. Asif, J. Tang, and S. Harrer, "GraspNet: An efficient convolutional neural network for real-time grasp detection for low-powered devices," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4875–4882.

[18] S. Yu, D.-H. Zhai, Y. Xia, H. Wu, and J. Liao, "SE-ResUNet: A novel robotic grasp detection method," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 5238–5245, Apr. 2022.

[19] S. Yu, D.-H. Zhai, and Y. Xia, "EGNet: Efficient robotic grasp detection network," *IEEE Trans. Ind. Electron.*, vol. 70, no. 4, pp. 4058–4067, Apr. 2023.

[20] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, "Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 3, pp. 1384–1394, Jun. 2023.

[21] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[22] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5049–5059.

[23] K. Sohn et al., "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 596–608.

[24] Z. Xu et al., "Anti-interference from noisy labels: Mean-teacher-assisted confident learning for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 11, pp. 3062–3073, Nov. 2022.

[25] J. Yin et al., "Semi-supervised 3D object detection with proficient teachers," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 727–743.

[26] G. Zhou, D. Wang, Y. Yan, H. Chen, and Q. Chen, "Semi-supervised 6D object pose estimation without using real annotations," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5163–5174, Aug. 2022.

[27] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.

[28] L. Zhang, Y. Shi, Z. Shi, K. Ma, and C. Bao, "Task-oriented feature distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 14759–14771.

[29] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 9626–9633.

[30] Z. Zhou et al., "Local observation based reactive temporal logic planning of human-robot systems," *IEEE Trans. Automat. Sci. Eng.*, early access, Aug. 25, 2023, doi: 10.1109/TASE.2023.3304842.

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

[32] Y. Wu, F. Zhang, and Y. Fu, "Real-time robotic multigrasp detection using anchor-free fully convolutional grasp detector," *IEEE Trans. Ind. Electron.*, vol. 69, no. 12, pp. 13171–13181, Dec. 2022.

[33] S. Kumra, S. Joshi, and F. Sahin, "GR-ConvNet v2: A real-time multi-grasp detection network for robotic grasping," *Sensors*, vol. 22, no. 16, 2022, Art. no. 6208.

[34] R. Xu, F.-J. Chu, and P. A. Vela, "Gknet: Grasp keypoint network for grasp candidates detection," *Int. J. Robot. Res.*, vol. 41, no. 4, pp. 361–389, 2022.

[35] Y. Wu, Y. Fu, and S. Wang, "Information-theoretic exploration for adaptive robotic grasping in clutter based on real-time pixel-level grasp detection," *IEEE Trans. Ind. Electron.*, vol. 71, no. 3, pp. 2683–2693, Mar. 2024.

[36] P. Shukla, V. Kushwaha, and G. C. Nandi, "Development of a robust cascaded architecture for intelligent robot grasping using limited labelled data," *Mach. Vis. Appl.*, vol. 34, no. 6, 2023, Art. no. 99.