IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2019, accepted March 25, 2019, date of publication April 1, 2019, date of current version April 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908281

Multi-Object Grasping Detection With Hierarchical Feature Fusion

GUANGBIN WU[1,2], WEISHAN CHEN[1], HUI CHENG[3], WANGMENG ZUO[4], (Senior Member, IEEE), DAVID ZHANG[5], (Fellow, IEEE), AND JANE YOU[2]

[1]State Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin 150001, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3]School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[4]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
[5]School of Science and Engineering, The Chinese University of Hong Kong at Shenzhen, Shenzhen 518172, China

Corresponding author: Weishan Chen (cws@hit.edu.cn)

This work was supported in part by the GRF Fund from the HKSAR Government, in part by the Central Fund from The Hong Kong Polytechnic University, in part by the Major Program of Science and Technology Planning Project of Guangdong Province under Grant 2017B010116003, in part by the NSFC Fund under Grant 61332011 and Grant 61671182, and in part by the Shenzhen Fundamental Research Fund under Grant JCYJ20150403161923528.

ABSTRACT Grasping in cluttered and tight scenes is a necessary skill for intelligent robotics to achieve more general application. Such universal robotics can use their perception abilities to visually identify grasps from a stack of objects. However, most existing grasping detection methods based on deep learning just focus on estimating grasping pose with single-layer features. In this paper, we present a novel grasp detection algorithm termed as multi-object grasping detection network, which can utilize hierarchical features to learn object detector and grasping pose estimator simultaneously. The network is mainly composed of two branches: 1) Object detection branch which is based on the single shot multibox detection approach to discriminate object categories and locate object positions by bounding boxes; 2) Grasping pose estimation branch where hierarchical features are fused together to predict grasping position and orientation. To improve grasping detection performance, attention mechanism is employed in hierarchical feature fusion. For evaluating the proposed model, we build a multi-object grasping dataset where every image contains numerous different graspable objects. The extensive experiments demonstrate that the multi-object grasping detection method achieves the state-of-the-art performance on both object detection and grasping pose estimation.

INDEX TERMS Deep learning, object detection, pose estimation, robotic grasping, hierarchical feature fusion.

I. INTRODUCTION

Grasping is a fundamental capability for intelligent robots to accomplish many kinds of autonomous manipulation. Although being a very simple and intuitive action for human beings, visual-based grasping tasks are still a challenging problem for intelligent robots so far, such as background noise, variations in viewpoints and scene complexity [1]. Hence, it is quite essential to develop an effective approach to detect grasping pose for an identical object in unstructured environments.

Previous algorithms for the robotic grasping pose perceiving are mainly based on model-matching tasks. For example, in order to optimize the robotic grasping quality, Gallegos *et al.* [2] provided a grasp search algorithm with higher dimensional continuation tools to satisfy contact constraints between a robotic hand and an object. Pokorny *et al.* [3] firstly defined spaces of graspable objects, then mapped new objects to these spaces to discover grasps. Despite these works are powerful, they rely on complete and accurate three-dimensional (3D) models, which would not be feasible and effective when an intelligent robot is interacting with unstructured environments.

To take advantage of RGBD sensors, it is more convenient to capture 2D images than reconstruct the 3D model to detect grasping configurations. Lenz *et al.* [4] demonstrated that five-dimensional grasp configuration in two-dimensional space can be projected to the three-dimensional

The associate editor coordinating the review of this manuscript and approving it for publication was Xiwang Dong.

space to guide robotic grasping. The grasp configuration in three-dimensional space has been simplified to the seven-dimensional representation by Jiang *et al.* [5]. Besides, deep convolutional networks can extract hierarchical features automatically and have achieved great success in object detection [13], classification [27], scene understanding [28] and tracking [29]. Taking 2D images as input, many researches based on deep learning are booming up recently and have yielded many outstanding fruits.

The most common deep learning approach for 2D robotic grasp prediction is a sliding window detection framework [4], [6], [7]. In the framework, many patches are extracted from an image and then fed into the deep neural network in sequence to predict a grasp pose for each patch. Finally, the prediction grasping pose with the highest detection confidence is selected as the final prediction results. However, this method is computationally expensive and drastically decreases the speed of the task.

To avoid sliding windows, end-to-end approaches were employed in [8]–[10]. In these studies, one image passing through the network once can obtain the detection results directly. These approaches start with the strong prior that every image contains a single grasp target. As for in the real-world environments cluttered with many objects, Guo *et al.* [12] utilized the segment method to isolate objects to predict their grasp poses. Asif *et al.* [11] firstly used the Faster R-CNN to locate object bounding boxes and then estimated object grasp poses. Although these methods achieve some outstanding performance on grasping detection, they cannot detect objects and estimate object grasping poses simultaneously.

To solve grasping challenges in unstructured environments, we propose a novel grasping detection approach based on deep neural network. The proposed approach can exploit hierarchical features to learn object detector and grasping pose estimator simultaneously in an end-to-end manner, which mainly consists of two branches. One exploits multiple feature maps to predict object categories and object locations represented by bounding boxes in an image. This branch is based on single shot multibox detector method [13]. Another branch is used to estimate robotic grasping configurations via hierarchical feature fusion. In order to take advantage of hierarchical features effectively and suppress noise, an attention mechanism is exploited while features extracted from different layers of deep neural network are fused. Besides, to reduce parameter redundancy and increase computational efficiency, these two branches share a backbone in the neural network.

Since most previous works just focus on studying grasping pose estimation for an image piece with one object but not take object detection into consideration, they can use the Cornell Grasping Dataset to evaluate the performance of their approaches. In an image of Cornell Grasping Dataset, there is only one object with a clean background, which does not satisfy our requirements. In order to evaluate the multi-object grasping detection model, we build a Multi-object Grasping Dataset, which is composed of six categories of objects. And in every image, there are different instances from one to ten. On the extensive experimental evaluations, the proposed method achieves outstanding performance on both object detection and grasping pose estimation.

The main contributions of this paper can be summarized as follows.

1) A multi-object grasping detection model based on deep neural network is proposed, which can detect objects and estimate grasping poses simultaneously in an end-to-end manner. With the grasping detection algorithm, intelligent robots can classify cluttered objects automatically in the real world by grasping manipulation.

2) To improve object grasping detection performance, hierarchical features fused by an attention mechanism are exploited in the deep neural network.

3) We build a Multi-object Grasping Dataset where there are numerous object instances with different grasping configurations in each image. This dataset can be applied to validate numerous multi-object grasping detection algorithms for intelligent robotics manipulation.

4) Utilizing the Multi-object Grasping Dataset, we validate the proposed approach which can achieve outstanding performance on both object detection and grasping pose estimation.

The remainder of this paper is organized as follows: related works about robotic grasping and object detection are reviewed in Section II. Section III represents the problem formulation of the multi-object grasping detection. Section IV demonstrates the multi-object grasping detection algorithm in detail. Section V discusses the experimental results on both Multi-object Grasping Dataset and Cornell Grasping Dataset. Section VI ends this paper by providing several concluding remarks.

II. RELATED WORKS

In this section, we firstly review the robotic grasping detection approaches, then introduce some outstanding object detection algorithms with deep learning.

A. ROBOTIC GRASPING

Grasping is an unremarkable behavior for human beings in our daily life. However, it is a challenging problem in the area of intelligent robotics. The exploration on how to enable the robot to grasp objects in unstructured cluttered environments dexterously and intelligently as humans remains an active research topic.

Previous works studying robotic grasping mainly focus on 3D model matching approaches. "GraspIt!" [14] is one of the presentative works, which is a robotic grasping simulator and widely exploited in the robotic grasping community. Many kinds of robotic hands and 3D object models imported into this simulator can create arbitrary 3D projections of the 6D grasp wrench space. For further studying, Goldfeder *et al.* [15] released the Columbia Grasp Database with a large number of 3D models of graspable objects. This grasping database can be utilized in "GraspIt!" and

acts as a benchmark for robotic grasping tasks. Besides, Gallegos *et al.* [2] studied the grasping pose and the desired contact points with 3D models of the objects. However, these model-based methods are only applicable when the precise 3D models of the objects are known in advance, which is usually not true when a robot works in a new environment. It is more of a theoretical than a practical method to solve the robotic grasping problem.

Compared with obtaining the precise 3D models of objects, it is more convenient to get the visual information of objects from varies kinds of visual sensors. Moreover, Lenz *et al.* [4] showed that five-dimensional grasp representation in 2D space can be projected back to a seven-dimensional representation in 3D space to guide robotic grasping. By this reduction in dimension, the computational efficiency is improved greatly. Consequently, alternative robotic grasping approaches based on image detection are proposed recently. Saxena *et al.* [16] primitively utilized the probabilistic model to predict grasping pose of an object from a 2D image. For obtaining a good grasp in 2D space, Le *et al.* [17] detected the object grasping pose with multiple contact points. These previous algorithms predict object grasping region using the hand-designed features, which are highly depended on the expert knowledge and the raw data.

Whereas, deep learning has a powerful ability to learn the distinguished features of images automatically and has yielded many outstanding achievements on classification [33]–[35], object detection [20]–[22] and tracking [30]–[32]. Inspired by these achievements, many researchers have been carrying out many works in the field of robotic grasping detection. A remarkable deep grasping detection work is from Lenz *et al.* [4], which generates numerous rectangle candidates with different sizes, location and orientation to feed into deep neural network for classification. The candidate with the highest classification score is taken as the final grasp detection result. Unfortunately, the sliding window approach to predict grasping pose with so many image patches is time-consuming. To increase the grasping detection speed, Redmon and Angelova [8] exploited a single-stage regression method which made images pass through the networks once to predict the grasping bounding box directly. Park *et al.* [18] proposed a classification based grasping detection algorithm with multiple-stage spatial transformer networks. The networks firstly cropped image patches with certain location and orientation and then fed into grasping classifier to judge the grasping confidence. Finally, the best grasp pose was selected using the max pooling.

For object grasping in cluttered environments, Dogar *et al.* [19] firstly moved away the obstacles in the desired path using the robotic hand and then grasped the target objects. However, this physics-based approach may damage the target objects brutally if the calibration is not precise. Afterwards, [12] segmented objects from unstructured scene and then predicted grasping configuration separately. Asif *et al.* [11] employed object detection method to predict the bounding boxes of target objects and then estimated the grasping pose for each image patches. Although these works can predict object pose in cluttered scene, they take two or more steps.

In order to make up for the insufficient of above algorithms, a novel grasping detection approach is proposed in this paper, which can utilize hierarchical features to detect object categories and estimate grasping configurations simultaneously in an end-to-end manner. By this approach, every target object in unstructured and cluttered scene will be given a best grasp pose estimation even a complex image passes through the deep neural network only once.

B. OBJECT DETECTION

The aims of object detection are to identify the categories and locations of objects in a given scene. With the advantages of deep learning, object detection study has yielded many great achievements.

One of key components of deep object detection approaches is the region-based networks, such as regions with convolutional neural networks (R-CNN) [20], fast R-CNN [21] and faster R-CNN [22]. The original R-CNN [20] utilizes the sliding window method to detect categories and locations of objects in an image. As the heavy deep neural networks require to compute great quantities of image patches, it is time consuming and known to be extremely slow. For improving detection speed, fast R-CNN [21] employs the sliding window on a feature space rather than on an image. Although it can significantly decrease the computation cost, it still needs to process a great many candidates of object detection. Ren *et al.* [22] proposed the faster R-CNN based on region proposal network to generate the region proposals in an efficient and accurate way. By sharing convolutional features with the down-stream detection network, this method can greatly reduce computation time at the region proposal step. However, the region-based detection methods cannot detect objects by processing an image only once.

Recently, single shot architectures, such as you only look once (YOLO) [23] and single shot multibox detector (SSD) [13], have shown state-of-the-art performance both on the computing efficiency and the object detection precision. Instead of evaluating many object candidate windows, YOLO [23] exploits the whole topmost feature map to predict both confidence of multiple object categories and object locations by processing an image only once. SSD [13] further develops the object detection approach based on YOLO. SSD discretizes the multi-scale convolutional bounding boxes to attach hierarchical feature maps at the highest layers of the neural networks. Compared with other object detection methods, this representation significantly reduces the computing time and improves the detection accuracy even with a smaller input image size. Under consideration of these advantages, we utilize the SSD approach in our work for object detection.

III. PROBLEM FORMULATION

Given an input image I containing numerous different objects, the multi-object grasping detection model is required

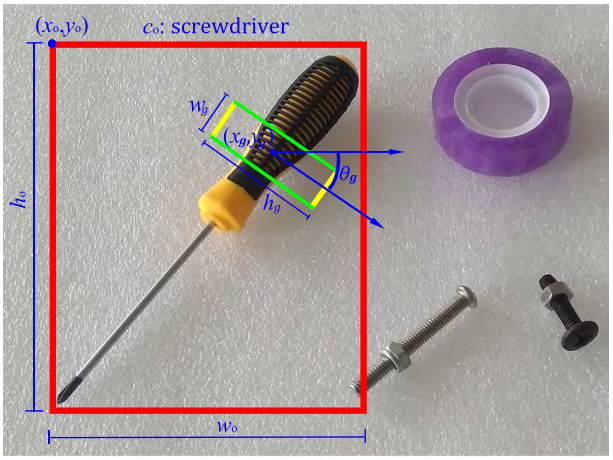


FIGURE 1. Representations for object detection and grasping configuration.

to learn the detection representation O and grasping configuration G for each object. Just as described by Lenz *et al.* [4], a five-dimensional grasping configuration can be projected into a seven-dimensional configuration for robotic grasping in a real scene. In this paper, we focus on studying five-dimensional grasping representations using 2D images for a parallel-plate gripper of a robot. As the Fig. 1 shows, a grasping pose in a 2D image can be presented as follows.

$$G = \{x_g, y_g, h_g, w_g, \theta_g\}$$

where (x_g, y_g) represents the centroid of the grasping rectangle. The term h_g and w_g stand for the height and width of the grasping rectangle respectively. And θ_g is the grasping orientation given by the angle between the grasping height h_g and the horizontal axis of the image.

Object detection representation of an object shown in the Fig. 1 is described as:

$$O = \{x_o, y_o, h_o, w_o, c_o\}$$

where (x_o, y_o) stands for the minimum vertex coordinate of an object bounding box. h_o and w_o are height and width of the object bounding box. c_o is the object detection confidence.

IV. PROPOSED METHOD

Fig. 2 shows the overall architecture of the multi-tasks deep neural network for robotic grasping detection. The proposed architecture is designed based on VGG16 [24] and consists of two main branches: i) Object detection branch which is utilized to predict the categories and locations of objects; ii) Grasping pose estimation branch which fuses hierarchical deep features by attention mechanism to discriminate robotic grasping pose. In the final layer, these two branches are coalesced to calculate an ideal robotic grasping configuration for every object in a cluttered scene.

A. OBJECT DETECTION

Given an input image $I \in R^{W\times H\times C}$, the object detection branch can formulate the representation $O = \{x_o, y_o, h_o, w_o, c_o\}$ for every object. The terms W and H, C respectively stand for the width, the height and the channel number of the image. While a RGB image is exploited as input data, C equals to 3.

In this paper, object detection part is based on the single shot multibox detector (SDD300) [13]. And the backbone of deep neural network is the VGG16 network. As the Fig. 2 describes, multiple features utilized to predict both bounding boxes and location confidence of objects are extracted from Conv4-3, Conv7, Conv8-2, Conv9-2, Conv10-2 and Conv11-2 layers of the network.

As the SSD method describes, the object detection objective loss is composed of localization loss and the classification confidence loss shown by the function 1.

$$L_o(\delta_o, c_o, l_o, g_o) = \frac{1}{N}(L_{o\_conf}(\delta_o, c_o) + \alpha L_{o\_loc}(\delta_o, l_o, g_o)) \tag{1}$$

where N denotes the number of matched default boxes, α is the weight parameter which is set to 1 in this paper similar to [13].

The classification confidence loss is the solftmax loss described as follows.

$$L_{o\_conf}(\delta_o, c_o) = -\sum_{i\in Pos}^{N} \delta^p_{o\_ij} \log(\hat{c}^p_{o\_i}) - \sum_{i\in Neg} \log(\hat{c}^0_{o\_i}) \tag{2}$$

where $\hat{c}^p_{o\_i} = \frac{\exp(c^p_{o\_i})}{\sum_p \exp(c^p_{o\_i})}$; $\delta^p_{o\_ij} \in \{0, 1\}$ denotes an indicator which matches the *i-th* default box to the *j-th* ground truth box of category p; $\hat{c}$ stands for the classification confidence.

For learning object location in an image, the $smooth_{L1}$ loss is applied as the localization loss shown in the function 3.

$$L_{o\_loc}(\delta_o, l_o, g_o) = \sum\sum \delta^k_{o\_ij} smooth_{L1}(l^m_{o\_i} - \hat{g}^m_{o\_j}) \tag{3}$$

where

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 0.5 \\ |x| - 0.5 & otherwise \end{cases} \tag{4}$$

$$\hat{g}^{x_o}_{o\_j} = (g^{x_o}_{o\_j} - d^{x_o}_i)/d^{w_o}_i, \quad \hat{g}^{y_o}_j = (g^{y_o}_j - d^{y_o}_i)/d^{h_o}_i$$
$$\hat{g}^{w_o}_{o\_j} = \log(g^{w_o}_{o\_j}/d^{w_o}_i), \quad \hat{g}^{h_o}_{o\_j} = \log(g^{h_o}_{o\_j}/d^{h_o}_i) \tag{5}$$

where $m \in \{x_o, y_o, w_o, h_o\}$; $(d^{x_o}_i, d^{y_o}_i)$ is the center coordinate of the *i-th* grid cell which corresponds to the *j-th* ground truth bounding box g_j; $d^{w_o}_i$ and $d^{h_o}_i$ indicate the width and the height of the grid cell respectively; $\hat{g}$ and l represent the offsets of the ground truth bounding box and the prediction bounding box.

B. GRASPING POSE ESTIMATION

Instead of directly learning the grasping poses in 3D space, we learn the grasping configurations utilizing 2D images and then project them into the real scene with depth information to guide robotic grasping. In the 2D space, an oriented grasping rectangle can be represented by the five-dimensional

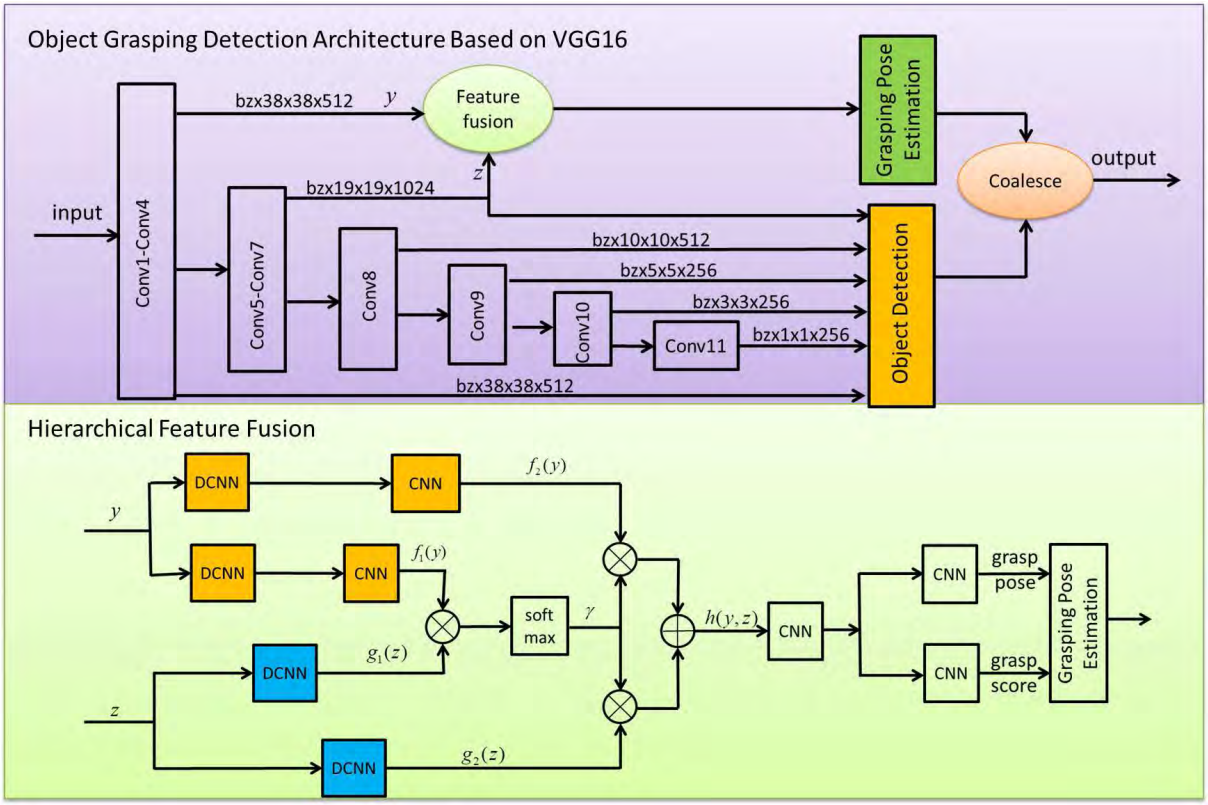


FIGURE 2. The architecture of multi-object grasping detection. Overall framework of the proposed model is shown in the first row which mainly consists of two branches: One is to detect object categories and locations with multi-layer features extracted from Conv4-3, Conv7, Conv8-2, Conv9-2, Conv10-2 and Conv11-2 layers; Another branch (reseda part) is to predict grasping configurations using hierarchical feature fusion approach detailed in the second row. Hierarchical features (y and z) coming from Conv4-3 and Conv7 layers are fused with attention mechanism. DCNN and CNN denote the deconvolutional and convolutional operations respectively. Finally, these two branches are coalesced to predict an ideal robotic grasping configuration for every object.

vector $G = \{x_g, y_g, h_g, w_g, \theta_g\}$. Inspired by SSD, we divide the input image into multiple grid cells and then predict a grasping configuration for each grid cell. In order to take full advantage of useful information and to suppress noise, hierarchical features fused by an attention mechanism are exploited in the proposed model.

1) GRASPING REPRESENTATION

To learn numerous robotic grasping representations in an image, we divide the image into 57 × 57 anchor boxes and make an assumption that there is only one grasping pose representation in each anchor box. Consequently, Multi-Box objective is employed to handle oriented grasping rectangle just as the SSD model. The objective function consists of two parts: the grasping location loss L_{g_loc} and the grasping confidence loss L_{g_conf} described by the function 6.

$$L_g(\delta_g, c_g, l_g, t_g) = \frac{1}{M}(L_{g\_conf}(\delta_g, c_g) + \beta L_{g\_loc}(\delta_g, l_g, t_g)) \tag{6}$$

where M is the number of the matched pose and negative grasping rectangles. $smooth_{L1}$ loss is employed as the grasping location loss for reducing the discrepancy between the predicted oriented rectangles (l_g) and the ground truth oriented rectangles (t_g). In the loss function, the grasp rectangle orientation is implied by the $sin\ 2\theta$ and $cos2\theta$. In other word, the grasping configuration can be represented by $G' = \{x_g, y_g, h_g, w_g, sin\ 2\theta, cos2\theta\}$, where $\theta = atan2(sin\ 2\theta, cos2\theta)$. The regression loss for the anchor offsets is described by the function 7.

$$L_{g\_loc}(\delta_g, l_g, t_g) = \sum_{i\in\{Pos,Neg\}}^{M} \sum_{m\in G'} \delta_{g\_ij} smooth_{L1}(l^m_{g\_j} - \hat{t}^m_{g\_i}) \tag{7}$$

where $\delta_{g\_ij} = \{0, 1\}$ is an indicator for matching the j-th predicted grasping rectangle to the i-th ground truth grasping rectangle. To be specific, if the two center points of predicted and ground truth grasping rectangles are in the same anchor boxes, $\delta_{g\_ij} = 1$, otherwise, $\delta_{g\_ij} = 0$. Besides,

$$\begin{aligned} \hat{t}^{x_g}_{g\_i} &= (t^{x_g}_{g\_i} - a^{x_g}_{g\_i})/a^{w_g}_{g\_i}, \quad \hat{t}^{y_g}_{g\_i} = (t^{y_g}_{g\_i} - a^{y_g}_{g\_i})/a^{h_g}_{g\_i} \\ \hat{t}^{w_g}_{g\_i} &= t^{w_g}_{g\_i}/a^{w_g}_{g\_i}, \quad \hat{t}^{h_g}_{g\_i} = t^{h_g}_{g\_i}/a^{h_g}_{g\_i} \\ \hat{t}^{sin2\theta}_{g\_i} &= sin2\theta, \quad \hat{t}^{cos2\theta}_{g\_i} = cos2\theta \end{aligned} \tag{8}$$

where $(a^{x_g}_{g\_i}, a^{x_g}_{g\_i})$ is the center points of the i-th ground truth grasping anchor a_{g_i}. $a^{w_g}_{g\_i}$ and $a^{h_g}_{g\_i}$ denote the width and height of the i-th anchor cell respectively.

The softmax loss is utilized in grasping pose classification given by the function 9.

$$L_{g\_conf}(x_g, c_g) = -\sum_{i\in\{Pos,Neg\}}^{M} \delta_{g\_ij} \log\left(\frac{e^{c^p_{g\_i}}}{\sum_p e^{c^p_{g\_i}}}\right) \quad (9)$$

where $p \in \{0, 1, 2\}$ indicates the grasping category. $p = 1$ stands for the negative grasping; $p = 2$ denotes the negative grasping, while $p = 0$, the grasping pose does not belong to these two.

In the paper, the parameter β used to weight the localization and confidence losses is set to 1.

2) HIERARCHICAL FEATURE FUSION

In order to take full advantage of hierarchical features, deep features extracted from Conv4-3 and Conv7 layers are fused with attention mechanism. The detailed hierarchical feature fusion is shown in the second row of Fig. 2, where DNN and CNN represent the deconvolutional and convolutional operations respectively.

Let $y \in R^{(W_1\times H_1)\times C_1}$ and $z \in R^{(W_2\times H_2)\times C_2}$ be the image features extracted from Conv4_3 and Conv7 layers respectively. For calculating the attention matrix, features y and z are firstly transformed into two feature spaces f_1 and q_1, where $f_1(y) = \Phi_{1f}(\Psi_{1f}(y))$, $q_1(z) = \Phi_{1q}(\Psi_{1q}(z))$. The symbols Φ and Ψ stand for deconvolutional and convolutional operations respectively. The gating coefficient matrix γ is indicated in Fig. 2, which is formulated as follows:

$$\gamma_{ji} = \frac{\exp(\mu_{ij})}{\sum_{i=1}^{N}\exp(\mu_{ij})}, \quad where\ \mu = f_1(y)^T q_1(z) \quad (10)$$

where γ_{ji} denotes the extent to which the model attends to the *i-th* location when synthesizing the *j-th* region. Then the output of the attention fusion layers $h(y, z) = \left(h_1(y, z), \cdots, h_j(y, z), \cdots, h_{N_h}(y, z)\right)^T \in R^{N_h\times C_h}$ is given as function 11, where $N_h = W_h \times C_h$.

$$h_j(y, z) = \sum_{i=1}^{N} \gamma_{ji} f_{2i}(y) + \omega_1 \sum_{i=1}^{N} \gamma_{ji} q_{2i}(z) \quad (11)$$

where $f_2(y) = \Phi_{2f}(\Psi_{2f}(y))$, $q_2(z) = \Phi_{2q}(\Psi_{2q}(z))$. ω_1 is a learnable weighted parameter, initialized as 1.0.

In above mathematical expressions, W_i, H_i and C_i denote the width, the height and the channel number of image features respectively, where $i = 1, 2, h$.

After hierarchical feature fusion with attention mechanism, $h(y, z)$ passes through two convolutional layers to predict grasping configurations and corresponding grasping scores.

C. GLOBAL OBJECTIVE FUNCTION

Combining the loss functions of object detection and grasping pose estimation, we yield the global objective function to optimize the parameters Θ of multi-object grasping detection network. The global objective function is described by the function 12.

$$L_G(\Theta) = L_o(\delta_o, c_o, l_o, g_o) + \xi L_g(\delta_g, c_g, l_g, t_g)) \quad (12)$$

where ξ is a penalty parameter utilized to weight the grasping loss, which is set to 0.5 in our paper.

In back propagation, we exploit the stochastic gradient decent (SGD) approach to learn the model. Hence, the parameters Θ updated in epoch $v+1$ can be described as following.

$$\Theta^{v+1} \leftarrow \Theta^{v} - \eta \frac{\partial L_G(\Theta)}{\partial \Theta} \quad (13)$$

where η stands for learning rate.

D. COALESCENCE

After hierarchical feature fusion with the attention mechanism, grasping poses are estimated for all grid cells of an image. Subsequently, the best grasp configuration for each object can be calculated according to grasping confidence scores and object detection. The maximum grasping scores of all predicted objects are formulated by the function 14 in the coalescent step.

$$S_{og} = \bigcup_{i=1}^{N} max\left(\bigcup_{j=1}^{M} (\delta(A_j \cap A_i) \cdot s_{gj}\right) \quad (14)$$

where

$$\delta(A_j \cap A_i) = \begin{cases} 1, & A_j \cap A_i > 0 \\ 0, & otherwise \end{cases} \quad (15)$$

N and M stand for the number of the predicted objects and grasping grid cells in an image. s_{gj} is the grasping score in *j-th* grid cell. A_j and A_i indicate the bounding box area of the *j-th* detection object and the area of the *i-th* grid cell respectively.

With the highest grasping confidence score of every object, all grasping detections $D_{og} = \{D_{og,i}(G_{og,i}, O_{og,i})\}_{i=1}^{N}$ are calculated after the combined prediction processing, where $G_{og,i} = \{x_{og,i}, y_{og,i}, h_{og,i}, w_{og,i}, \theta_{og,i}, s_{og,i}\}$ and $O_{og,i} = \{x_{_{og,i}}, y_{_{og,i}}, h_{_{og,i}}, w_{_{og,i}}, c_{_{og,i}}\}$ are grasping representation and object detection results of the *i-th* predicted object.

V. EXPERIMENTS AND DISCUSSION

To evaluate the effectiveness and efficiency of the multi-object grasping detection model, we did some experiments on both Multi-object Grasping Dataset and Cornell Grasping Dataset [4]. Meanwhile, we analyzed the effectiveness of each component in the hierarchical feature fusion layers.

A. DATASETS AND EVALUATION METRIC

1) CORNELL DATASET

The Cornell Grasping Dataset is composed of 885 images with 244 different objects. In each image, there is only one object labeled with multiple ground truth positive and negative grasping configurations. However, these objects are without category labels and location bounding boxes. In order to evaluate the proposed method on this dataset, we assign the category labels for all objects to be 1 and acquire their rectangle bounding boxes manually to be the ground truth information.

TABLE 1. Object instances distributed in each class.

Class	Dice	Lego	Screw	Screwdriver	Pen	Tape	Total
Instance	3	6	8	4	5	6	32

2) MULTI-OBJECT GRASPING DATASET

Since there is no public grasping dataset for multi-object grasping detection studying in the cluttered scene, we build a Multi-object Grasping Dataset to evaluate our proposed algorithm. The dataset consists of 1022 images with 1-10 different object instances in each image. Moreover, there are totally 6 classes with 3-8 object instances in each class detailedly shown in the Table 1. Similar to Cornell Grasping Dataset, the ground truth grasping configurations of each object are composed of numerous positive and negative grasps.

3) EVALUATION METRIC

In this paper, we focus on studying object grasping detection for robots with parallel grippers. There are two evaluation metrics for such robotic grasping: point metric and rectangle metric. The point metric [16], [36] evaluates a grasp with distances between the predicted point and all actual grasp centers. If any of these distances less than a specified threshold, the grasp is considered to be successful. However, the metric does not take grasping orientation into consideration, which might overestimate the performance of an algorithm for robotic applications [4].

The rectangle metric proposed by Jiang *et al.* [5] evaluates robotic grasps by both grasping angle and Jaccard index criteria simultaneously. A candidate grasp configuration is considered to be valid while it satisfies these two criteria detailed as follows.

1) The angle difference between ground truth grasp G_t and predicted grasp G_p is less than 30°. It is because that a robot with parallel gripper can grasp an object successfully even the predicted angle error is large.

2) The Jaccard index of the ground truth grasp G_t and predicted grasp G_p is greater than 25%, e.g.

$$J(R_t, R_p) = \frac{|R_t \cap R_p|}{|R_t \cup R_p|} \tag{16}$$

where R_t is the area of the ground truth grasping rectangle, and R_p is the area of the predicted grasping rectangle. According to the definition, the Jaccard index is the same as the interest of region (IoU) threshold in object detection study, which includes grasping center point and gripper open wide information. Since the ground truth rectangle can define a large space of graspable rectangle [4], [5], the predicted rectangle overlapped by 25% with one of ground truth grasps is still a good grasp while its orientation is correct. Besides, Multi-object Grasping Dataset is built under the Cornell Grasping Dataset criterion. Therefore, similar to previous works [4], [5], [9]–[11], we evaluate our model utilizing the rectangle metric.

B. RESULTS AND DISCUSSIONS

While evaluating the proposed algorithm, VGG16 [24] is implemented as the backbone of the network. We randomly initialize weights in all new layers with a zero-mean Gaussian distribution and standard deviation 0.03. Similar to SSD object detection method [13], other layers are initialized by pre-trained model on the ImageNet [25]. Then we fine-tune deep networks using stochastic gradient descent (SGD) algorithm with learning rate 0.001, momentum 0.9, weight decay 0.0005 and batch size 16. For satisfying the SSD300 object detection model, all images are re-sized to 300 × 300 pixels with re-scaled values −1.0-1.0. Full training and testing codes are built on Tensorflow.

1) EXPERIMENTS ON MULTI-OBJECT GRASPING DATASET

In the experiments, we randomly select 80% images as training data and 20% images as test data. There are 3527 object instances, 29096 positive and 39617 negative grasps in training dataset and 1173 object instances, 9617 positive and 13418 negative grasps in test dataset respectively, which are detailed described in Table 2.

TABLE 2. Details of training and test data in multi-object grasping dataset.

	Image	Instance	Positive Grasp	Negative Grasp
Training	767	3527	29096	39617
Test	255	1173	9617	13418
Total	1022	4700	38713	53035

TABLE 3. Detection results on multi-object grasping dataset using the backbone of VGG16 net.

Method	Object Detection Accuracy (%)	Grasping Pose Estimation Accuracy (%)	Speed fps
MGD_SF19	93.34	85.12	33.35
MGD_SF38	94.67	85.91	32.53
MGD_HF	95.31	86.41	32.14
MGD_HFA	95.31	87.10	31.42

Experimental results are shown in Table 3, where MGD_HFA is the proposed multi_object grasping detection model whose hierarchical features are fused via attention mechanism. Others are taken as baseline models which are also indicated by the fusion part in the first row of the Fig. 2. The MGD_SF19 and MGD_SF38 are multi_object grasping detection approaches using single layer features whose dimensions are 19 × 19 × 1024 and 38 × 38 × 512 respectively. The features z and y are separately extracted from Conv7 and Conv4_3 layers shown in Fig. 2. MGD_HF is the multi_object grasping detection with simple hierarchical feature fusion. Both models of MGD_SF and MGD_HF are detailed in the Fig. 3.

The first two columns in Table 3 describe the prediction accuracies of object detection and grasping pose respectively. Compared with single feature (SF) models, hierarchical feature fusion approaches (HF) yield outstanding performance

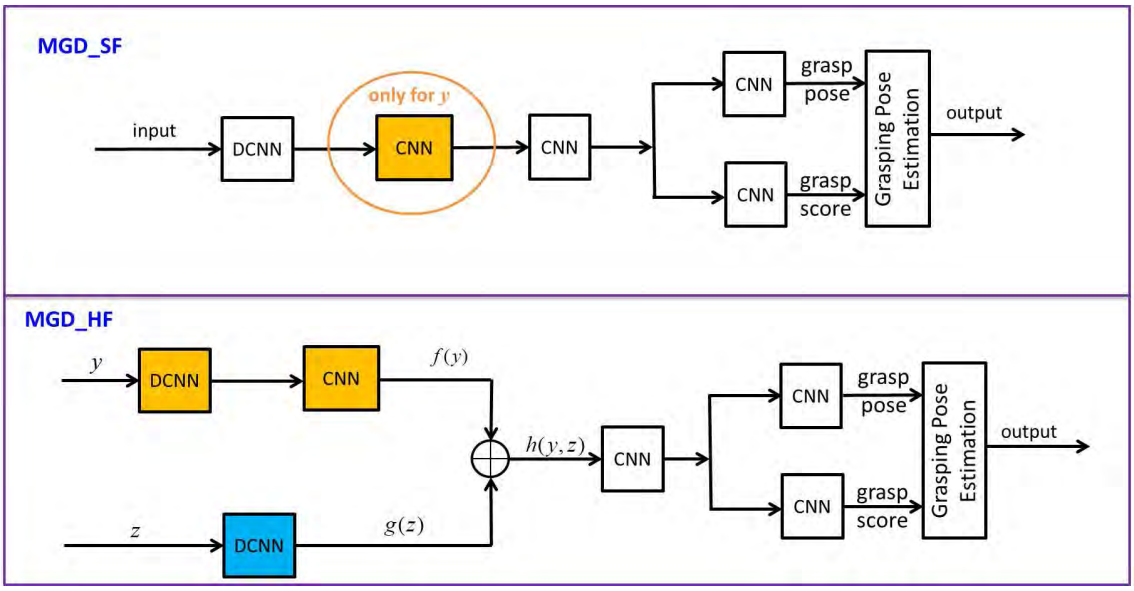


FIGURE 3. Architectures of MGD_SF and MGD_HF. The input of MGD_SF is the features *y* coming from Conv4-3 layer for MGD_SF38, or the features *z* coming from Conv7 for MGD_SF19. Compared with the MGD_SF19, MGD_SF38 has more one CNN layer given by the orange block in the image.

on the prediction accuracy. Furthermore, the grasping prediction accuracy of MGD_HFA 87.10% is higher than that of the simple MGD_HF 86.40%. For further studying the computing time, time-consuming experiments are conducted with the batch size 1 on a device equipped with GTX 1080Ti GPU, Intel Xeon W-2133@3.6GHz CPU, 16G memory. As the last column shows in Table 3, the prediction speed of the proposed model is 31.42 frames per second (fps), about 31.83 milliseconds per image, which meets the requirements of real-time applications (30 fps [23]). Therefore, the hierarchical features fusion strategy employed in our model can achieve more accurate grasping detection performance without too much time burden.

To make comparison with other object detection algorithms, we conducted more experiments with YOLO-based and faster-RCNN-based object grasping detection methods. Analogous to SSD-based multi-object grasping detection, the VGG16 is utilized to be the backbone network, and hierarchical features extracted from Conv4-3 and Conv7 are fused using attention mechanism to predict grasping configurations. Object detection parts are the same as the original works [22], [23]. Experimental results on our dataset are shown in Table 4, where YMGD_HFA and FMGD_HFA stand for YOLO-based and faster-RCNN-based multi-object grasping detection methods respectively. As the Table 4 shows, both accuracy and speed performance of the two methods are inferior to the SSD-based method, especially the detection speeds which are 15.54 fps for YMGD_HFA and 5.80 fps for FMGD_HFA, much slower than 31.42 fps for MGD_HFA.

Besides, we validate the universality of the proposed method employing the ResNet50 [37] and DenseNet121 [38] as backbone networks. In the ResNest50-based architecture,

TABLE 4. Comparison methods on multi-object grasping dataset.

Method	Object Detection	Grasping Pose Estimation	Speed
	Accuracy (%)		fps
YMGD_HFA	87.67	82.94	15.54
FMGD_HFA	93.86	85.29	5.80
DMGD_HFA	97.25	88.54	25.31
RMGD_HFA	96.57	87.83	28.24
MGD_HFA	95.31	87.10	31.42

multi-features coming from conv3_4, conv4_6, conv5_3 and last layers are used to detect objects, and hierarchical features extracted from conv3_4 and conv4_6 layers are fused via attention mechanism to estimate grasping configurations. For the DenseNet121-based network, multi-features coming from dense block2, dense block3, dense block4 and last layer are employed to detect objects; hierarchical features extracted from dense block2 and dense block3 are exploited to estimate grasping configurations with hierarchical feature fusion. Experimental parameters and learning strategies of the two models are the same as those of the VGG16-based method. Experimental results are shown in Table 4, where DMGD_HFA and RMGD_HFA denote the proposed methods based on DenseNet121 and Resnet50 respectively. As the Table 4 shows, the two multi-object grasping detection approaches with DenseNet121 and ResNet50 backbone networks achieve outstanding performance on both accuracy and efficiency of the multi-object grasping detection, which implies that the proposed method can generalize to other backbone networks. However, the prediction speeds of DenseNet121-based and ResNet50-based algorithms are slower than that of VGG16-based method since these networks have much more learned parameters.

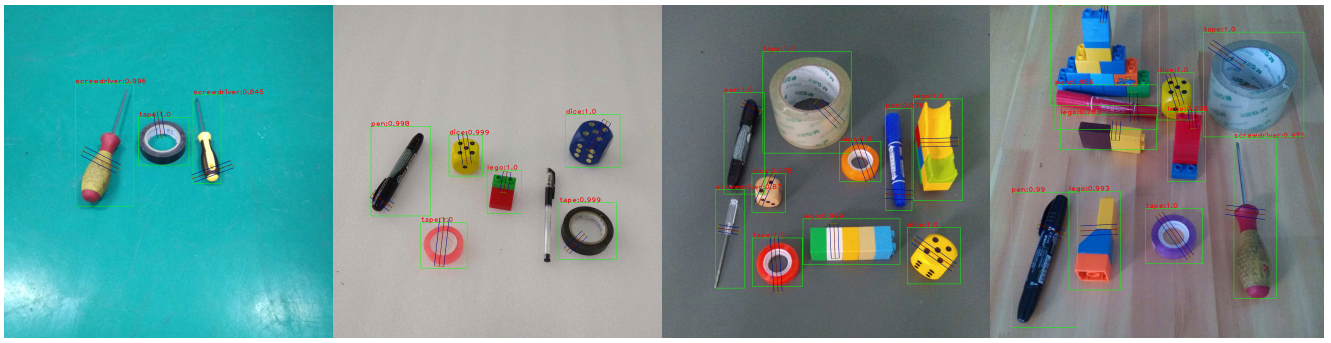


FIGURE 4. Grasping detection results on our dataset.

2) EXPERIMENTS ON CORNELL GRASPING DATASET

To make a comparison with the existing state-of-the-art methods [4], [5], [7], [8], [26], we extensively evaluate the proposed method on the Cornell Grasping Dataset. In this grasping dataset, objects are without category labels and location bounding boxes. In order to satisfy our model, all objects are assigned labels with 1 (named object) and bounded rectangle boxes manually. In the experiments, RGB images of the Cornell Grasping Dataset are employed to detect grasping configurations.

TABLE 5. Grasping detection performance on Cornell Grasping Dataset. "—" means that the corresponding method cannot discriminate object.

Method	Object Detection	GP Estimation	Speed
	Accuracy (%)		fps
Fast Search [5]	—	63.8	0.02
SAE [4]	—	73.9	0.07
Direct_Regression [8]	—	84.4	13.16
Two_stage_close_loop [7]	—	85.3	7.10
Jacquard [26]	—	86.88	—
MGD_SF19	97.74	84.16	39.60
MGD_SF38	97.83	86.42	39.20
MGD_HF	98.04	86.43	37.52
MGD_HFA	98.19	86.88	36.75

Similar to the experiments on Multi-object Grasping Dataset, 80% images of Cornell Grasping Dataset are randomly selected as training data and remains are test data. The experimental results are described in Table 5. Compared with multi_object grasping detection models, such as MGD_HF, MGD_SF19 and MGD_SF38, the MGD_HFA approach obtains the highest prediction accuracies on both object detection and grasping pose estimation which are 98.19% and 86.88% respectively. Although the computation efficiency of the proposed method is slightly lower than that of other multi-object grasping models, its prediction speed reaches to 36.75 fps which meets the demand for most real-time detection tasks.

Fast search [5], SAE [4], direct regression [8], two_stage close loop [7] and Jacquard [26] are state-of-the-art grasping detection approaches. Employing the multiple modal features (RGBD information), these methods only can predict the grasping configurations, whose prediction performances are shown in the Table 5. Compared with these methods, the proposed model cannot only detect objects and predict grasping configurations simultaneously, but also achieve the outstanding performances on both efficiency and accuracy for multi-object grasping detection. Especially, the prediction speed reaches to 36.75 fps which is over 1837 times faster than that of fast search method. Hence, the MGD_HFA is much more suitable for generalizing to complex scene for intelligent robotic grasping classification.

TABLE 6. Comparison methods on Cornell Grasping Dataset.

Method	Object Detection	Grasping Pose Estimation	Speed
	Accuracy (%)		fps
YMGD_HFA	90.13	81.72	17.46
FMGD_HFA	96.45	85.31	6.09
DMGD_HFA	99.02	87.92	25.83
RMGD_HFA	98.81	87.43	30.34
MGD_HFA	98.19	86.88	36.75

We also compare the proposed method with YMGD_HFA and FMGD_HFA on Cornell Grasping Dataset. Network parameters and learning strategies of the two models are the same as those of experimental evaluation on Multi-object Grasping Dataset. Experimental results are shown in Table 6. As conclusions are drawn on Multi-object Grasping Dataset, prediction performance of the both YMGD_HFA and FMGD_HFA methods is inferior to those of MGD_HFA, especially the prediction speed. Therefore, the SSD-based approach can contribute much more to the multi-object grasping detection.

In addition, the ResNet50-based and Densenet121-based architectures (RMGD_HFA and DMGD_HFA) are exploited to validate the universality of the proposed method on Cornell Grasping Dataset. Their experimental sets are similar to those on Multi-object Grasping Dataset. Experimental results are shown in Table 6, which indicate that both RMGD_HFA and DMGD_HFA models yield outstanding performance on both object detection and grasping pose estimation. Especially, DMGD_HFA achieves the highest detection accuracy: 99.02% for object detection and 87.92% for grasping pose estimation. Prediction speeds are 25.83 fps for DMGD_HFA

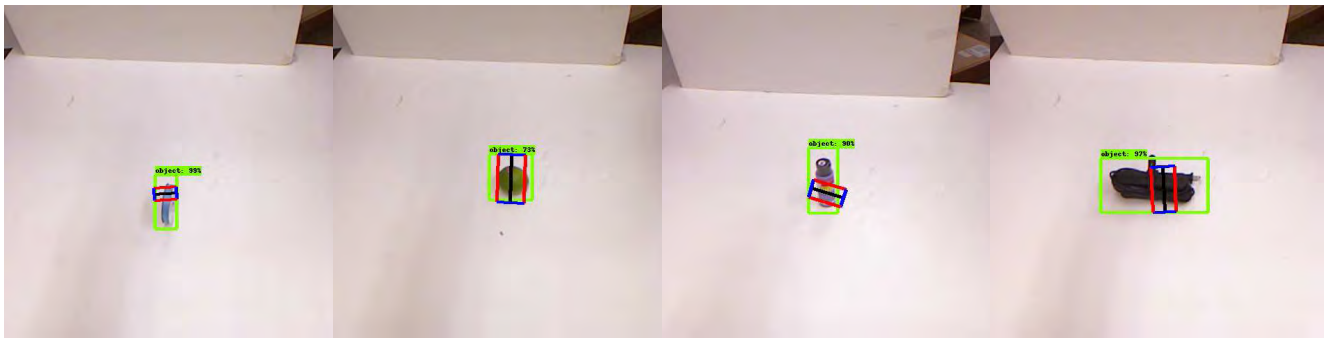


FIGURE 5. Grasping detection results on Cornell Grasping Dataset.

and 30.34 fps for RMGD_HFA, which are slower than 36.75 fps for the VGG16-based methods (MGD_HFA). It is mainly because that there are a larger number of parameters in ResNet50 and Densenet121. According to these experimental results, we can observe that the proposed method can obtain a significant performance gain regardless of the use of backbone structures.

VI. CONCLUSION

In this paper, we proposed a novel multi-object grasping detection approach which can detect objects and predict grasping configurations simultaneously in clustered environments. The proposed model mainly consists of two branches: object detection branch based on SSD and grasping pose estimation branch. The coalescent approach is employed in the last layer for calculating the final multi-object grasps. In order to take full advantage of useful information extracted from the image and suppress noise, hierarchical features fused by the attention mechanism are utilized in the grasping pose estimation branch, which can improve the grasping detection performance. As for experimental evaluations, the proposed model with hierarchical feature fusion achieves the state-of-the-art performance on the efficiency and accuracy of object grasping detection on both Multi-object Grasping Dataset and Cornell Grasping Dataset.

ACKNOWLEDGMENT

The authors would like to thank NVIDIA Corporation providing the Tesla K40c GPU for our research.

REFERENCES

[1] U. Asif, J. Tang, and S. Harrer, "Ensemblenet: Improving grasp detection using an ensemble of convolutional neural networks," in *Proc. British Mach. Vis. Conf. (BMVC)*, Newcastle, U.K., Sep. 2018, pp. 10–21.

[2] C. R. Gallegos, J. M. Porta, and L. Ros, "Global optimization of robotic-grasps," in *Proc. 7th Robot. Sci. Syst.* Los Angeles, CA, USA: University of Southern California, 2011, pp. 289–296.

[3] F. T. Pokorny, K. Hang, and D. Kragic, "Grasp moduli spaces," in *Proc. 9th Robot. Sci. Syst.* Berlin, Germany: Technische Universität Berlin, Jun. 2013.

[4] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, 2015.

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

[6] J. Wei, H. Liu, G. Yan, and F. Sun, "Robotic grasping recognition using multi-modal deep extreme learning machine," *Multidimensional Syst. Signal Process.*, vol. 28, no. 3, pp. 817–833, Jul. 2017.

[7] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Adv. Mech. Eng.*, vol. 8, no. 9, pp. 1–12, 2016.

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

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

[10] J. Watson, J. Hughes, and F. Iida, "Real-world, real-time robotic grasping with convolutional neural networks," in *Towards Autonomous Robotic Systems*. New York, NY, USA: Springer, 2017, pp. 617–626.

[11] U. Asif, J. Tang, and S. Harrer, "Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices," in *Proc. IJCAI*, Jul. 2018, pp. 4875–4882.

[12] D. Guo, T. Kong, F. Sun, and H. Liu, "Object discovery and grasp detection with a shared convolutional neural network," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 2038–2043.

[13] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[14] A. T. Miller and P. K. Allen, "Graspit! A versatile simulator for robotic grasping," *IEEE Robot. Automat. Mag.*, vol. 11, no. 4, pp. 110–122, Jun. 2004.

[15] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen, "The columbia grasp database," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2009, pp. 1710–1716.

[16] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[17] Q. V. Le, D. Kamm, A. F. Kara, A. Y. Ng, "Learning to grasp objects with multiple contact points," in *Proc. IEEE Int. Conf. Robot. Automat.*, May 2010, pp. 5062–5069.

[18] D. Park and S. Y. Chun. (2016). "Classification based grasp detection using spatial transformer network." [Online]. Available: https://arxiv.org/abs/1803.01356

[19] M. Dogar, K. Hsiao, M. Ciocarlie, and S. Srinivasa, *Physics-Based Grasp Planning Through Clutter*. Cambridge, MA, USA: MIT Press, 2018.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[21] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, May 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, May 2016, pp. 779–788.

[24] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

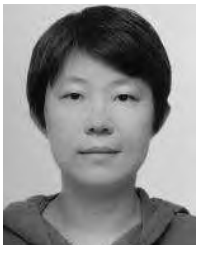
[25] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[26] A. Depierre, E. Dellandréa, and L. Chen. (2016). "Jacquard: A large scale dataset for robotic grasp detection." [Online]. Available: https://arxiv.org/abs/1803.11469?context=cs

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1–4.
[28] S. Di, H. Zhang, C.-G. Li, X. Mei, D. Prokhorov, and H. Ling, "Cross-domain traffic scene understanding: A dense correspondence-based transfer learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 745–757, Sep. 2018.
[29] S.-H. Bae and K.-J. Yoon, "Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking," *IEEE Trans. pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 595–610, Apr. 2018.
[30] X. Lu, H. Huo, T. Fang, and H. Zhang, "Learning deconvolutional network for object tracking," *IEEE Access*, vol. 6, pp. 18032–18041, 2018.
[31] E. Gundogdu and A. A. Alatan, "Good features to correlate for visual tracking," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2526–2540, Mar. 2018.
[32] S. Yun, J. Choi, Y. Yoo, K. Yun, and J. Y. Choi, "Action-driven visual object tracking with deep reinforcement learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2239–2252, Jun. 2018.
[33] P. Wang, F. Ye, X. Chen, and Y. Qian, "Datanet: Deep learning based encrypted network traffic classification in sdn home gateway," *IEEE Access*, vol. 6, pp. 55380–55391, 2018.
[34] Y. Deng, Z. Ren, Y. Kong, F. Bao, and Q. Dai, "A hierarchical fused fuzzy deep neural network for data classification," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 1006–1012, Mar. 2017.
[35] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Jun. 2017.
[36] A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng, "Robotic grasping of novel objects," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1209–1216.
[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2016, pp. 770–778.
[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2017, pp. 4700–4708.

GUANGBIN WU received the B.S. degree in mechanical manufacturing and automation from Yantai University, in 2009, and the M.S. degree in marine engineering from Ningbo University, China, in 2012. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Robotics and System, Harbin Institute of Technology. His research interests include domain adaptation, object detection, and intelligent robot.

WEISHAN CHEN received the B.E. and M.E. degrees in precision instrumentation engineering and the Ph.D. degree in mechatronics engineering from the Harbin Institute of Technology, China, in 1986, 1989, and 1997, respectively. Since 1999, he has been a Professor with the School of Mechatronics Engineering, Harbin Institute of Technology. His research interests include ultrasonic driving, smart materials and structures, and bio-robotics.



HUI CHENG received the B.Eng. degree in electrical engineering from Yanshan University, Qinhuangdao, China, the M.Phil. degree in electrical and electronic engineering from The Hong Kong University of Science and Technology, and the Ph.D. degree in electrical and electronic engineering from The University of Hong Kong. She was a Postdoctoral Fellow with The Chinese University of Hong Kong, from 2006 to 2007. She is currently an Associate Professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou. Her current research interests include intelligent robots and networked control systems.

WANGMENG ZUO received the Ph.D. degree in computer application technology from the Harbin Institute of Technology, Harbin, China, in 2007, where he is currently a Professor with the School of Computer Science and Technology. His current research interests include image enhancement and restoration, object detection, visual tracking, and image classification. He has published more than 70 papers in top-tier academic journals and conferences. He has served as a Tutorial Organizer in ECCV 2016, an Associate Editor of the *IET Biometrics* and the *Journal of Electronic Imaging*, and the Guest Editor of *Neurocomputing*, *Pattern Recognition*, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

DAVID ZHANG (F'08) received the bachelor's degree in computer science from Peking University, Beijing, China, the M.Sc. and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1982 and 1985, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 1994.

From 1986 to 1988, he was a Postdoctoral Fellow with Tsinghua University, Beijing, and then as an Associate Professor with the Academia Sinica, Beijing. Since 2005, he has been a Chair Professor with The Hong Kong Polytechnic University, Hong Kong, where he is also the Founding Director of the Biometrics Research Centre (UGC/CRC) supported by the Hong Kong SAR Government, in 1998. He is currently a Professor with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China. He also serves as a Visiting Chair Professor with Tsinghua University and an Adjunct Professor with Peking University, Shanghai Jiao Tong University, Shanghai, China, HIT, and the University of Waterloo. He has authored or coauthored about 20 monographs and over 400 international journal papers. He holds around 40 patents from USA/Japan/Hong Kong/China.

Dr. Zhang is a Croucher Senior Research Fellow, a Distinguished Speaker of the IEEE Computer Society, and a fellow of IAPR. He is the Founder and an Editor-in-Chief of the *International Journal of Image and Graphics*, the Founder and a Series Editor of the International Series on Biometrics (KISB) (Springer), an Organizer of the International Conference on Biometrics Authentication, and an Associate Editor of more than ten international journals, including the IEEE TRANSACTIONS and so on. He was selected as a Highly Cited Researcher in Engineering by Thomson Reuters, from 2014 to 2017.

JANE YOU received the B.Eng. degree in electronic engineering from Xi'an Jiaotong University, in 1986, and the Ph.D. degree in computer science from La Trobe University, Australia, in 1992. She is currently a Professor with the Department of Computing, The Hong Kong Polytechnic University. She has been a Principal Investigator for two ITF projects (Innovation Technology Fund), four GRF projects (General Research Fund) supported by the Hong Kong Government and many other joint grants since she joined PolyU, in late 1998. So far, she has more than 280 research papers published. Her research output has led to three U.S. patents, technology transfers, and international awards. Her current research interests include image processing, medical imaging, computer-aided detection/diagnosis, and pattern recognition. She is also an Associate Editor of *Pattern Recognition* and other journals.