IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2020, accepted July 4, 2020, date of publication July 13, 2020, date of current version July 24, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3008763*

# Grasping Objects Mixed With Towels

**XIAOMAN WANG[1], XIN JIANG[1], (Member, IEEE), JIE ZHAO[1], SHENGFAN WANG[1], AND YUN-HUI LIU[1,2], (Fellow, IEEE)**

[1]Department of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen 518055, China
[2]T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong

Corresponding author: Xin Jiang (x.jiang@ieee.org)

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1309300, and in part by the Shenzhen and Hong Kong Joint Innovation Project under Grant SGLH20161209145252406.

**ABSTRACT** In logistics warehouse sorting, rubbish classification, and household services, scenarios exist in which rigid and soft objects are randomly piled together. In such situations, two major challenges arise in robotic picking tasks: the first is to distinguish rigid objects from soft objects, and the second is to grasp one object of each type at a time. In this study, we propose a novel robotic picking methodology for the grasping of objects mixed with towels. The proposed approach is based on a novel object detection method that can identify a rigid object placed in different directions using a rotational bounding box. Rigid objects can be separated from the mixed scene without object segmentation. Moreover, the grasping pose of a rigid object can be generated directly along its principal axis, without using a CAD model or specific pose detection method. The gripper opening width is determined according to the object size. Therefore, our method can detect whether other objects, particularly soft ones, exist around a rigid object. If no suitable grasping pose is available for the rigid objects, the grasping pose on a wrinkle of the towel is selected. The experiments demonstrate that our method can accomplish the picking task in scenes with mixed rigid and soft objects, thereby indicating its significance in robotic object detection and sorting.

**INDEX TERMS** Object detection, rotational bounding box, grasping pose, mixed objects, towels.

## I. INTRODUCTION

In terms of logistics warehouse sorting, rubbish classification, and household services, humans constantly need to deal with object sorting in mixed and disordered scenes. People, including children, can easily pick up a rigid object without pulling up its surrounding objects. Similarly, they can successfully grasp a soft object without catching the objects covered thereby, and can place the objects in the correct locations according to their categories. In contrast, it is rather difficult for a robot to accomplish such tasks. Robots demonstrate a high probability of picking both rigid and soft objects together, as illustrated in Fig. 1. In this study, we focus on object grasping when rigid objects are mixed with towels.

The critical problem to be solved is to distinguish the rigid objects from the towels. Because a towel is highly deformable, it is adaptive to the shape of the object that is placed therein. Thus, in a scene with mixed objects, it is difficult to separate the object from the towels by segmentation, particularly when the colors of the towel and object are similar. Therefore, an object detection method is used to locate the region of the rigid object, which also achieves the aim of separating the rigid object from the towels. In recent years, numerous object detection methods have been proposed, including Faster R-CNN [1], YOLO [2]–[4], and SSD [5]. The bounding box of the network output can be used to locate an object. However, object orientation is not considered in bounding boxes, which leads to detection results with multiple objects included in one bounding box.

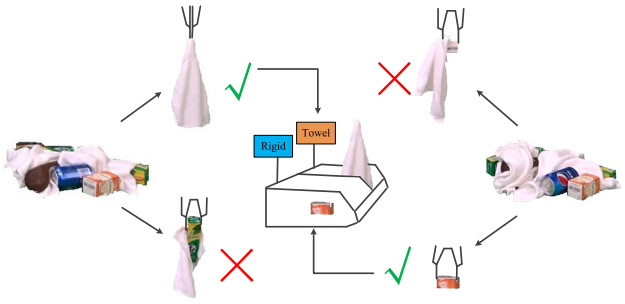



**FIGURE 1.** Overview of picking objects from situation of rigid objects mixed with towels.

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

Moreover, RGB images are used in these methods. If the color of the object is similar to the background, the detection results are unsatisfactory. In this study, inspired by the rotational bounding box introduced in [6], we extend YOLOV3 [4] and use RGB-D images as the input. In this manner, the network will output the rotational bounding boxes of the detected objects, so as to solve the problem that multiple objects are contained in one bounding box, which is convenient for the subsequent object grasping.

It has been established that, when planning a robotic grasping for a soft object, it is necessary to employ a pose detection method that differs from that for a rigid object, and vice versa. Although various impressive techniques have been achieved in the field of grasping, they have mainly focused on scenes with only one type of target [7]–[10]. When rigid and soft objects are mixed together, new challenges inevitably arise in grasping the objects. In this study, we select towels, which are in common use in daily life, as an example of soft objects for the problem of grasping mixed rigid and soft objects. One challenge is that the objects may be partly or totally covered by the towels. Another challenge is the selection of an appropriate grasping pose such that there is only one type of object within the gripper.

Upon comprehensive consideration of the above challenges, a novel grasping method based on object detection with a rotational bounding box is introduced in our work. Objects placed with different orientations can be detected by extending the YOLOv3 [4] network. Rigid objects can be separated from complex backgrounds without object segmentation. Different grasping methods are automatically selected according to the various object types. When grasping rigid objects, according to the detected information, a series of grasping poses can be generated directly along the principal axis without using a CAD model or a specific pose detection method. When grasping towels, we use the improved method of our previous work [11], which selects a grasping pose on the wrinkles of towels. This strategy can realize that only one type of object exists within the gripper. The experimental results demonstrate that the objects can be grasped successfully in scenes with mixed objects and placed in the designated box. The main contributions of this study are as follows:

- Based on the object detection results, we introduce an algorithm to accomplish the efficient grasping of objects mixed with towels.
- We propose the use of *RIoU* instead of the traditional *IoU* to measure the accuracy of the prediction bounding box, which can display the angle change between two rotational bounding boxes in an improved manner.
- We consider *GRIoU* as the loss of the rotational bounding box regression, which can improve the object detection performance.

## II. RELATED WORK

Robots may grasp multiple objects simultaneously in logistics warehouse sorting, rubbish classification, and household services. However, existing works have mainly focused on the scenario of grasping an object from a pile of rigid or soft objects. A detailed review of the previous studies on object detection and grasping is presented in the following sections.

### A. OBJECT DETECTION

In recent years, numerous methods for object detection have been proposed. However, the bounding boxes used in these studies are horizontally placed rectangles, consisting of four coordinate parameters, such as those in YOLOv3 [4] and SSD [5]. Moreover, the bounding boxes obtained by the above are generally substantially larger than the object. In particular, the detection results are significantly degraded when the object is not placed horizontally or vertically. These shortcomings of the horizontally placed rectangles make them unsuitable for the robotic object sorting strategy, namely, detection first followed by grasping. In the field of remote sensing image detection, to overcome this limitation, Liu *et al.* [6] proposed a new detection technique that can output an additional angle parameter in comparison with the conventional method. Thus, the object orientation angle can be determined. On the basis of the Faster R-CNN, Xu *et al.* [12] proposed that, by learning the offset of four points on the non-rotation rectangle, a quadrilateral can be formed to detect multi-oriented objects.

### B. GRASPING POSE DETERMINATION FOR RIGID OBJECTS

Gualtieri *et al.* [13] and ten Pas *et al.* [14] presented an impressive method for determining the candidate grasping pose by means of point clouds. With reference to the study of [14] and the classic work of PointNet [15], Liang *et al.* [16] introduced a new deep learning method known as PointNetGPD. The advantage of this method is that the spatial geometry of the point cloud in the graspable region can be understood better. Based on depth images, Mahler *et al.* [17] proposed an outstanding network known as GQ-CNN for grasping pose determination. More importantly, a relatively effective grasping pose for most daily objects can also be obtained using their method. Moreover, substantial research has been conducted on grasping pose determination, specifically for the Cornell grasping dataset [18]. Furthermore, multiple grasping of rectangular boxes can be generated on the image by [10], [19], and [20].

### C. GRASPING POSE DETERMINATION FOR SOFT OBJECTS

At present, the works on grasping soft objects, such as cloths and towels, focus on manipulation, including folding, hanging, and classification. In these studies, the premise is successful object grasping. Because cloths and towels may be randomly deformed, the grasping pose determination technique is quite different from that for rigid objects. An appropriate candidate grasping pose can be determined according to the aim of picking up clothes and its present poses. As demonstrated in [21], any point can be selected as

the grasping position. However, this approach often results in air grasping or grasping the target object together with other objects. In the work of [9], the center of the segmented region was selected as the candidate grasping point. In [22] and [11], a point that was at the highest position and on a wrinkle of the clothes was selected as the grasping position, respectively. In the procedure of folding, corner points are often considered as the grasping positions [9]. For further information on grasping clothes, please refer to the survey in [23].

In summary, when grasping rigid objects, existing works have mainly focused on the generation of grasping poses on the objects. Moreover, in manipulation on soft objects, the grasping pose is generated in different positions of the soft objects according to the different tasks. The two grasping methods are very different. Therefore, in scenes with mixed objects, it is very important to distinguish the two object types and to select different grasping poses according to the various object types. Furthermore, the special properties in the grasping of mixed rigid objects and towels must also be considered. For example, a grasping pose may be suitable for a rigid object; however, when part of a towel is near the object or under it, when the gripper is closed, the towel will also be grasped. Therefore, it is almost impossible to complete the task by using the above methods alone or by means of a simple combination of A, B, and C. In this study, based on the results of our new object detection method, we introduce a novel methodology to attempt to overcome the above challenges.

## III. METHOD

### A. PROBLEM STATEMENT

In logistics warehouse sorting, rubbish classification, and household services, robots not only need to grasp an object of each type successfully, but also need to recognize the category of the objects. There are two main solutions to this problem: first grasping and then detecting the object, or conversely, first detecting and then grasping the object. The common problems are to distinguish a rigid object from a soft one and to achieve the grasping of mixed rigid and soft objects. In the first strategy, occlusion occurs when recognizing the object grasped by the hand. The second strategy also exhibits certain shortcomings. The traditional detection method always leads to the detection of more than one object in a bounding box, particularly, when items are mixed together. In this study, we propose a novel object detection method that can identify objects placed in different orientations with rotational boxes. The method can effectively solve the problem that multiple objects are contained in one bounding box. On this basis, not only can the object category be recognized, but a rigid object can also be separated from the other objects. Moreover, we propose a robotic picking methodology that can select different grasping methods according to various object types.

### B. RIoU AND LOSS OF ROTATIONAL BOUNDING BOX REGRESSION

The extended YOLOv3 [4] is adopted to detect objects placed in different orientations. In comparison with the current YOLOv3 [4], the extended version can output an additional angle to represent the orientation of an object. Furthermore, the previous anchor box is no longer applicable in the extended YOLOv3 [4]. Therefore, the rotational anchor box is introduced into the extended network, as illustrated in Fig. 2. The angles of the rotational anchor boxes used in our method are composed of nine values: 10°, 30°, 50°, 70°, 90°, 110°, 130°, 150°, and 170°.

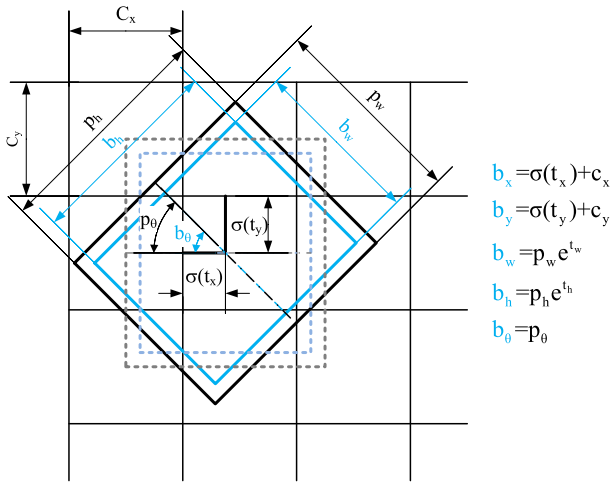



**FIGURE 2.** Rotational bounding boxes with dimension priors and location prediction. Apart from the additional angle information, the predicted parameters are exactly as in YOLOv3 [4]. Moreover, the network input is replaced with RGB-D, and *RIoU* is used to measure the overlap of two rotational bounding boxes.

After training, the parameters (including $(t_x, t_y, t_w, t_h, \theta)$) can be predicted. As described in YOLOv3 [4], the cell offsetting from the left upper corner of the image is $(c_x, c_y)$, and the size and angle of the bounding box prior can be expressed as $(p_w, p_h)$ and $p_\theta$, respectively. Thus, the predicted results can be expressed as

$$b_x = \sigma(t_x) + c_x \tag{1}$$
$$b_y = \sigma(t_y) + c_y \tag{2}$$
$$b_w = p_w e^{t_w} \tag{3}$$
$$b_h = p_h e^{t_h} \tag{4}$$
$$b_\theta = p_\theta, \tag{5}$$

where $\sigma(x)$ represents a sigmoid function and $(b_x, b_y, b_w, b_h, \theta)$ represent the center, size, and angle of the predicted box, respectively.

The value $IoU$ is used to measure the accuracy of the detection results. A higher value of $IoU$ indicates a more accurate prediction result. However, for the rotational bounding box, the current $IoU$ cannot effectively represent the angle change

of the prediction box, as indicated in Fig. 4. Liu *et al.* [6] proposed the angle-related *IoU* (*ArIoU*) to define the *IoU* relating to the angle. It is represented as follows:

$$ArIoU\ (A, B) = \frac{area\left(\hat{A} \cap B\right)}{area\left(\hat{A} \cup B\right)} \left|\cos\left(\theta_A - \theta_B\right)\right|, \quad (6)$$

where $\theta_A$ and $\theta_B$ represent the rotational angles of bounding boxes A and B, respectively. In this case, $\hat{A}$ and A have the same parameters, except for the angle, and the angle of $\hat{A}$ is equal to $\theta_B$.

However, the above definition exhibits its own defects. When the change in the two rotational angles is small, the value of *ArIoU* does not change significantly, as illustrated in Fig. 4. Therefore, in this study, we present another definition of *IoU* for the rotational bounding box: *RIoU*. It is more sensitive to a slight change in the angle and can demonstrate the change trend of the angle difference between the two intersecting rectangles. The value is given by:

$$RIoU\ (A, B) = \frac{area\left(A' \cap B'\right)}{area\left(A' \cup B'\right)} \left|1 - \sin\left(\theta_A - \theta_B\right)\right|, \quad (7)$$

where $\theta_A$ and $\theta_B$ represent the rotational angles of bounding boxes A and B, respectively. Furthermore, $A'$ and $B'$ are obtained from A and B through R1 and R2 transformations, as shown in Fig. 3. It can be observed from Fig. 4 that *RIoU* is more sensitive to small changes in the angle. If the predicted box and truth overlap completely, the value of *RIoU* will be 1. If the angle between the predicted box and truth is $\pi/2$, or no overlap exists between them, the value of *RIoU* will be 0.

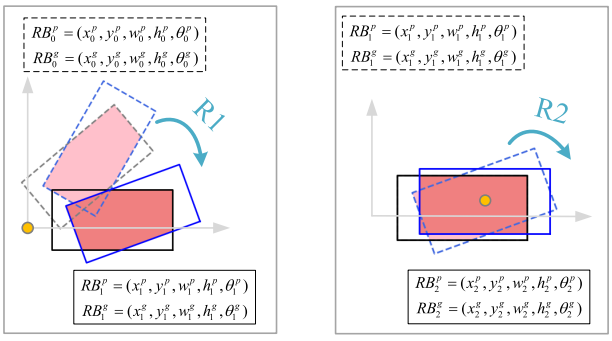
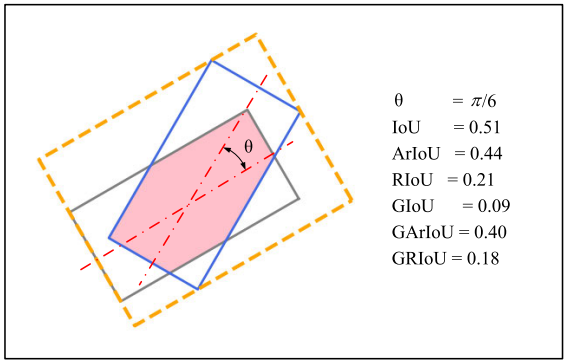
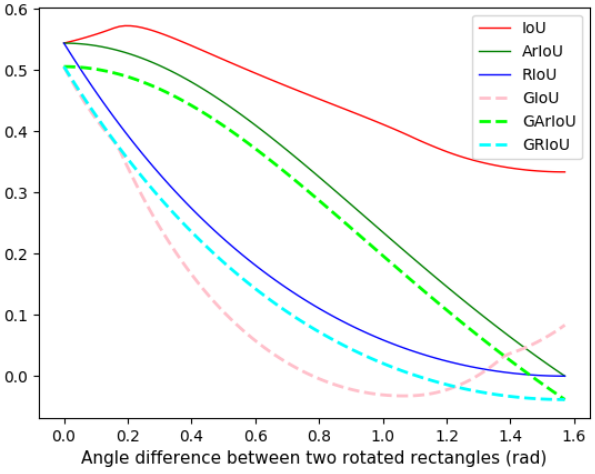



**FIGURE 3.** R1 and R2 transformations. Because it is challenging to compute the intersection of two rotational rectangles, the horizontal rectangles are obtained by R1 and R2 transformations. Following the R1 transformations, we obtain the horizontal rectangle and the relative positions of the two new rectangles are the same as before. Thereafter, the other rectangle is obtained by the R2 transformations.

When two rotational rectangles intersect, solving the intersection area is not as simple as determining the horizontal rectangle. To address the above problem, we transform two rotational rectangles ($A$ and $B$) of the initial position into two horizontal rectangles ($A'$ and $B'$) by means of two rotational transformations, as illustrated in Fig. 3. Following the R1 transformation, we obtain the horizontal rectangle $A1$ and




(a)




(b)

**FIGURE 4.** Comparison of *IoU* in different angle differences between two rotational rectangles in scene (a). The gray box remains in its current position and the blue box rotates around its center.

rotational rectangle $B1$. The rotational angle of $B1$ is the angle difference between $A$ and $B$. The intersection area of the two new rectangles is the same as the original intersection area. Moreover, the relative positions of the two new rectangles are the same as before. Thereafter, the horizontal rectangle $B'$ can be obtained by the R2 transformation. In this case, $B'$ and $B1$ have the same parameters, except that the rotational angle of $B'$ is 0, whereas $A'$ and $A1$ are exactly the same. As $A'$ and $B'$ are now horizontal rectangles, their intersection areas can easily be computed.

The *GIoU* [24] focuses on both the overlapping and non-overlapping areas of the two rectangles. It can provide a better reflection of the coincidence of the two rectangles. However, the *GIoU* [24] does not perform effectively in representing the change in the angle between the two rotational rectangles, as indicated in Fig. 4, although this concept is very useful for our work. Based on *GIoU*, we propose *GRIoU*, which can represent not only the change in the angle between two rotational rectangles, but also the coincidence between them, as illustrated in Fig. 4. Therefore, we use $L_{GRIoU}$ as the loss in the training process, as is presented in Algorithm 1.

**Algorithm 1** $L_{GArIoU}$ and $L_{GRIoU}$ as the Loss of Rotational Bounding Box Regression

**Input**: Predicted $RB_0^P$ and ground truth $RB_0^g$ rotational bounding box:
$RB^P = (x_0^p, y_0^p, w_0^p, h_0^p, \theta_0^p)$, $RB^g = (x_0^g, y_0^g, w_0^g, h_0^g, \theta_0^g)$.
**Output**: $RIoU$, $L_{GArIoU}$, $L_{GRIoU}$

1. $RB_1^P, RB_1^g$ is obtained by R1 transformation from $RB_0^P, RB_0^g$:
$RB_1^P = (x_1^p, y_1^p, w_1^p, h_1^p, \theta_1^p)$,
$x_1^p = x_0^p * cos(\theta_0^g) + y_0^p * sin(\theta_0^g)$,
$y_1^p = -x_0^p * sin(\theta_0^g) + y_0^p * cos(\theta_0^g)$,
$w_1^p = w_0^p, \quad h_1^p = h_0^p, \quad \theta_1^p = \theta_0^p - \theta_0^g$,
$RB_1^g = (x_1^g, y_1^g, w_1^g, h_1^g, \theta_1^g)$,
$x_1^g = x_0^g * cos(\theta_0^g) + y_0^g * sin(\theta_0^g)$,
$y_1^g = -x_0^g * sin(\theta_0^g) + y_0^g * cos(\theta_0^g)$,
$w_1^g = w_0^g, \quad h_1^g = h_0^g, \quad \theta_1^g = 0$.
2. $RB_2^P, RB_2^g$ are obtained by R2 transformation from $RB_1^P, RB_1^g$:
$RB_2^P = (x_2^p, y_2^p, w_2^p, h_2^p, \theta_2^p)$,
$x_2^p = x_1^p, y_2^p = y_1^p, w_2^p = w_1^p, h_2^p = h_1^p, \theta_1^p = 0$.
$RB_2^g = RB_1^g$.
3. Calcuating the are of $RB_0^P$,$RB_0^g$:
$A^p = w_0^p * h_0^p, \quad A^g = w_0^g * h_0^g$.
4. Calcuating the coordinates of the $RB_2^P, RB_2^g$:
$(x_{min}^P, y_{min}^P) = (x_1^P - w_1^P/2, y_1^P - h_1^P/2)$,
$(x_{max}^P, y_{max}^P) = (x_1^P + w_1^P/2, y_1^P + h_1^P/2)$,
$(x_{min}^g, y_{min}^g) = (x_1^g - w_1^g/2, y_1^g - h_1^g/2)$,
$(x_{max}^g, y_{max}^g) = (x_1^g + w_1^g/2, y_1^g + h_1^g/2)$.
5. Calcuating the intersection I of $RB_2^P, RB_2^g$:
$x_i^1 = max(x_{min}^P, x_{min}^g), \quad y_i^1 = max(y_{min}^P, y_{min}^g)$
$x_i^2 = min(x_{max}^P, x_{max}^g), \quad y_i^2 = min(y_{max}^P, y_{max}^g)$,
$I = max((x_i^2 - x_i^1), 0) * max((y_i^2 - y_i^1), 0)$.
6. Calcuating the smallest enclosing box $E$:
$x_e^1 = min(x_{min}^P, x_{min}^g), \quad y_e^1 = min(y_{min}^P, y_{min}^g)$
$x_e^2 = max(x_{max}^P, x_{max}^g), \quad y_e^2 = max(y_{max}^P, y_{max}^g)$,
$E = max((x_e^2 - x_e^1), 0) * max((y_e^2 - y_e^1), 0)$.
7. Calcuating the $ArIoU$, $RIoU$ of $RB_0^P, RB_0^g$:
$ArIoU = (A^p + A^g - I) * (cos(\theta_0^p - \theta_0^g))/(A^p + A^g)$.
$RIoU = (A^p + A^g - I) * (1 - sin(\theta_0^p - \theta_0^g))/(A^p + A^g)$.
8. Calcuating the $ArIoU$, $RIoU$ of $RB_0^P, RB_0^g$:
$GArIoU = ArIoU - (E - (A^p + A^g - I))/E$,
$GRIoU = RIoU - (E - (A^p + A^g - I))/E$.
9. $GArIoU$, $GRIoU$ as the loss of rotational bounding box regression:
$L_{GArIoU} = 1 - GArIoU$,
$L_{GRIoU} = 1 - GRIoU$.

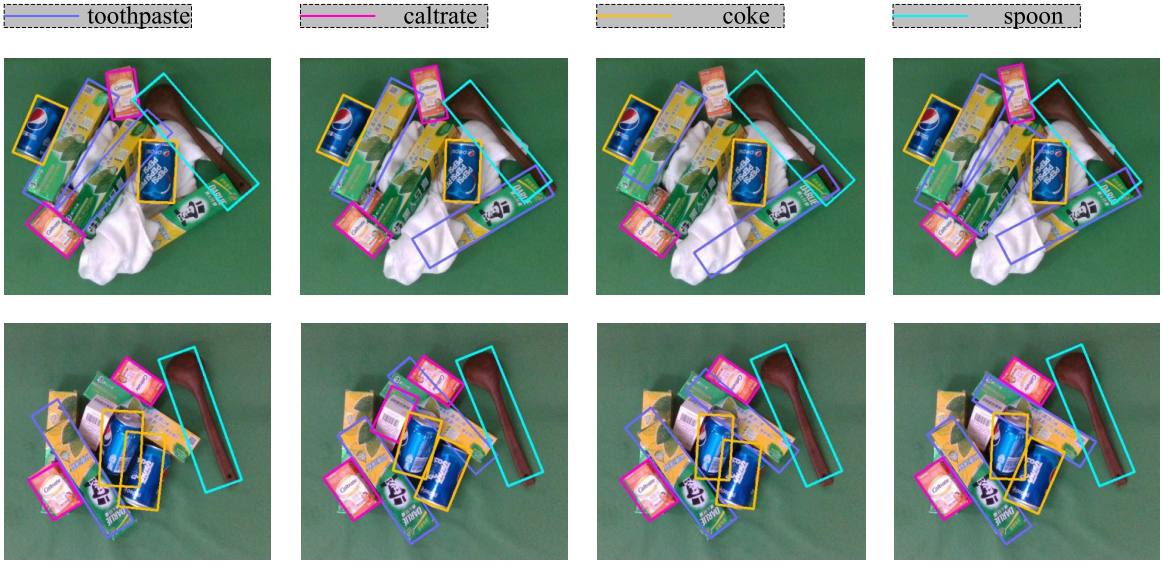
### C. GRASPING POSE DETERMINATION

#### 1) GRASPING POSE OF TOWELS

A towel is deformable and can be represented in thousands of states in space. Therefore, it is very difficult to separate towels with segmentation. In this study, we still select the grasping pose on wrinkles as in previous work [11]. Moreover, the point cloud used still does not contain color information. The advantage is that this method is applicable for towels of any color. In our work, the most convex point is not used as the grasping point, as in the previous study. In contrast, the center of the wrinkle is often regarded as the grasping point, which can be obtained by PCA (Principal Component Analysis) of the candidate wrinkle point cloud. This can prevent the grasping points from being located in an unfeasible grasping area of the wrinkle, while effectively reducing air grasping. Moreover, the grasping direction of the two-fingered gripper is along the principal axis of the wrinkle, and the principal axis can be computed by PCA. The grasping direction generated in this manner is generally perpendicular to the wrinkle. Consequently, grasping failure caused by an incorrect estimated grasping direction can effectively be avoided.

#### 2) GRASPING POSE OF RIGID OBJECTS

In a scene with mixed rigid objects and towels, owing to the shape adaptability of the towels to the rigid objects, the situation always exists in which the rigid objects may be surrounded and covered by towels. This requires the grasping pose to be generated by taking into account the above situation as far as possible. Therefore, it is necessary to generate the grasping poses along the principal axis of the object. By means of collision detection, whether there are towels or other objects near the object can be determined. Firstly, according to the detection results, the object can be selected as the candidate for grasping based on its score. Secondly, we map the 2D rotational bounding box to the 3D rotational bounding box, the benefit of which is that the rigid object can easily be separated from the mixed scene. Furthermore, this aids in generating the grasping pose along the principal axis of the object. Finally, through the PCA of the point cloud extracted from the 3D rotational bounding box, a series of grasping poses along the principal axis can be determined. The opening width of the two-fingered gripper reaches up to the dimension of the object. According to the PCA and size of the 3D rotational bounding box, we can obtain the center point $P_c(p_{c_x}, p_{c_y}, p_{c_z})$ of the extracted point cloud, the principal direction $(n_x, n_y, n_z)$, and the size $(w_r, h_r)$ of the object. Using this information, we can determine the grasping pose along the principal axis of the object. In this study, we sample three grasping points along the principal axis: $P_1$ $(p_{c_x}, p_{c_y}, p_{c_z})$, $P_2$ $(p_{c_x} - n_x * 0.3 * h_r, p_{c_y} - n_y * 0.3 * h_r, p_{c_z} - n_z * 0.3 * h_r)$, and $P_3$ $(p_{c_x} + n_x * 0.3 * h_r, p_{c_y} + n_y * 0.3 * h_r, p_{c_z} + n_z * 0.3 * h_r)$. Subsequently, we can identify




Iou: *ArIoU* [6] Loss: MSE[4] | Iou: *RIoU* (ours) Loss: MSE[4] | Iou: *ArIoU* [6] Loss: $L_{GArIoU}$ (ours) | Iou: *RIoU* (ours) Loss:$L_{RIoU}$(ours)

**FIGURE 5.** Detection results with rotational bounding box under different *IoU* and loss values.

which grasp pose may be suitable by collision detection. If the collision detection demonstrates that other objects exist around the object, that with the next-highest detection score is regarded as the subsequent candidate object.

#### 3) COLLISION DETECTION OF GRASPING POSE

A towel is adaptive to the two-fingered gripper and it is therefore unnecessary to conduct collision detection thereon, which is only required for the rigid objects in our work. Let $V(L) \subset R^3$ and $V(R) \subset R^3$ represent the volumes of the left and right fingers of a parallel-jaw gripper, respectively. Moreover, $V(L)$ and $V(R)$ represent the 3D rotational bounding boxes. Let $N \subset R$ represent the number of points in the 3D rotational bounding box. If $(N(V(L)) \leq C_T$ and $N(V(R)) \leq C_T)$, where $C_T$ is a constant value of 60, a collision-free grasping pose is indicated. Taking into account the noise from the camera, a constant $C_T$ is used in this work to increase the tolerance of the collision detection.

## IV. EXPERIMENTAL RESULTS

### A. DETECTION RESULTS

Because the color of the objects and the background may be similar, if only RGB is used as the input data of the network, it is difficult to distinguish between these. Therefore, we added an extra channel of depth to the network. Several detection results of the different *IoU* and loss values are presented in Table 1 and Fig. 5. Compared to *ArIoU*, *RIoU* could improve the detection performance with the same MSE loss as that used in YOLOV3 [4]. Moreover, the loss $L_{GRIoU}$ exhibited superior improvement. From Fig. 5, it is obvious that our method could predict the angles of the objects more accurately.

**TABLE 1.** Detection results with different IoU and loss.

| Loss / Evaluation | AP | AP75 |
|---|---|---|
| MSE[4](IoU: *ArIoU*[6]) | 70.106 | 88.010 |
| MSE[4](IoU: *RIoU*) | 73.574 | 91.336 |
| Relative improvement % | 4.9468% | 3.7791% |
| $L_{GArIoU}$ (IoU: *ArIoU*[6]) | 72.078 | 89.176 |
| Relative improvement % | 2.8129% | 1.3248% |
| $L_{GRIoU}$ (IoU: *RIoU*) | **76.732** | **93.808** |
| Relative improvement % | **9.4514%** | **6.5879%** |

### B. RIGID OBJECT GRASPING

The grasp pose of the rigid object and opening width ($1.5 * w_r$ of the object) of the two-fingered gripper could be determined based on the object detection information. The grasping pose occurred along the principal axis of the object, as illustrated in Figs. 6(c) and (e). Following collision detection, if one of the fingers was shown in red, this indicated that there were other objects near to it. Otherwise, the fingers were shown in blue, indicating that this was a proper grasping pose, as illustrated in Figs. 6(e) and (f). If the grasping poses of the candidate grasping object were all collisional, the object with the second-highest score became the candidate object, as indicated in Fig. 6(e). Following collision detection, the grasping pose was feasible for the third candidate object; thus, this grasping pose was selected. Moreover, for small-sized objects, collisions could occur in

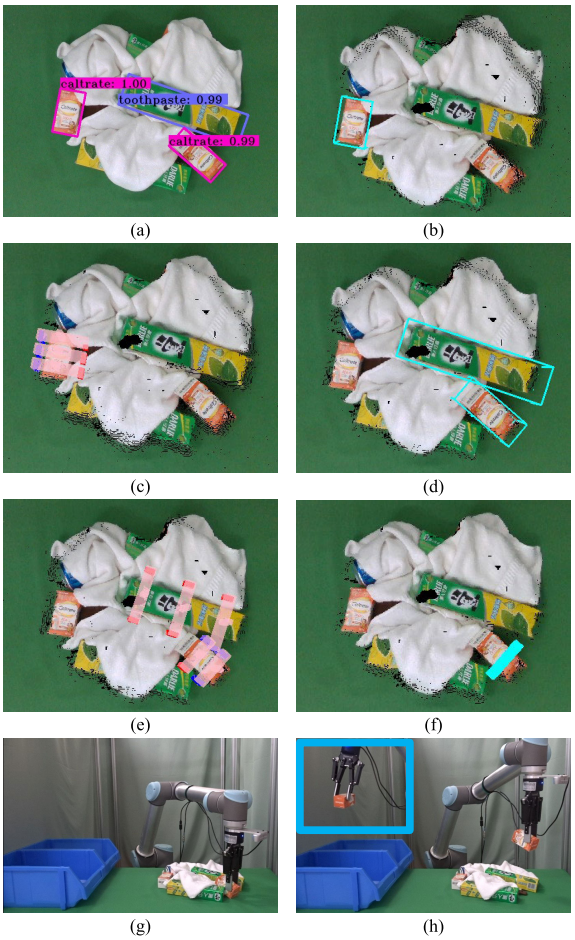



**FIGURE 6.** Entire procedure of picking rigid object. (a) Detection results. (b) Mapping 2D rotational bounding box to 3D for object with highest score. (c) Collision detection. The aims of (d) and (e) were the same as (b) and (c), respectively. (f) Obtaining feasible grasping pose. (g) and (h) represent the process of picking an object.

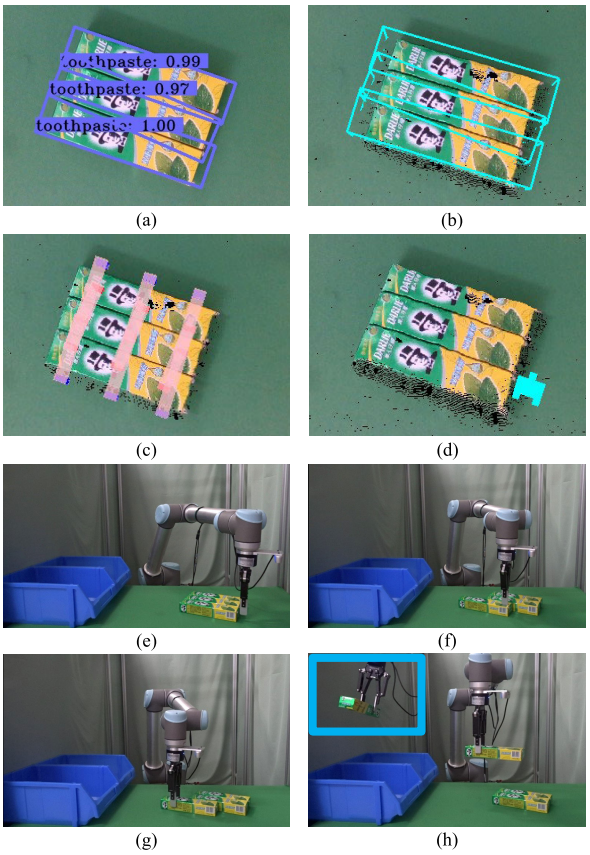



**FIGURE 7.** Entire procedure of pushing and grasping object. (a) Detection results. (b) Mapping 2D rotational bounding box to 3D for objects. (c) Collision detection. (d) Initial pushing pose. (e), (f), (g), and (h) represent the process of pushing and grasping an object.

all of the grasp poses along the principal axis direction, as illustrated in Fig. 8(c). The grasping direction could also be rotated by 90° and grasping could be attempted again, as indicated in Fig. 8(c). Furthermore, it had to be ensured that the width ($1.2 * h_r$ of the object) of the two-fingered gripper was no greater than the maximum opening limit.

When the objects were close together, it was difficult to grasp them successfully without the other actions, as illustrated in Fig. 7(a). In this case, we adopted the strategy of pushing and grasping to address the problem, as shown in Fig. 7. The pushed object was selected based on its score and the pushing direction was along the principal axis of the pushed object. The pushing end pose $P_e$ was set to $P_c$. The pushing initial pose was expressed as $P_i$ ($P_{i_x} = p_{c_x} + n_x * 0.5 * h_r, P_{i_y} = p_{c_y} + n_y * 0.5 * h_r, P_{i_z} = pc_z$).

### C. TOWEL GRASPING

Following the above grasping pose determination, if an appropriate grasping pose was still not detected, a towel would initially be grasped (see Fig. 8(c)). An overview of grasping a towel is presented in Fig. 8. Moreover, the opening distance of the two-fingered gripper was set to be slightly larger than the width of the wrinkle (in our experiments, the opening distance of the gripper was $g_w = 30$ mm). Moreover, it should be noted that the opening width of the two-fingered gripper jaw was very small and its end joints were passive joints. Therefore, even if there were objects under the towel, they would not be grasped.

In certain very special and rare cases, when the geometry of the covered objects is extremely clear, we can also detect these common geometries, such as rectangles and cylinders, as illustrated in Fig. 9. If the detected objects contained some regular shapes, as shown in Fig. 9(a), this indicated that certain objects were covered by towels. It was only necessary to exclude the region marked as unfeasible for grasping. Let $V\,(sob) \subset R^3$ represent the rotational bounding boxes for the specific objects (such as a cylinder). Let $V\,(rob) \subset R^3$ represent the rotational bounding boxes for the rigid objects (such as toothpaste and coke). Let $V'\,(rob) = 1.5 * 1.5 * \mathrm{V}(rob)$; that is, multiplying $(w_r, h_r)$ of $V\,(rob)$ by 1.5. Let $C\,(sob) \subset R^3$ represent the point cloud in $V\,(sob)$. Let $C\,(rob) \subset R^3$ represent the point cloud in $V\,(rob)$. Because several deviations could exist between the predicted and actual angles,

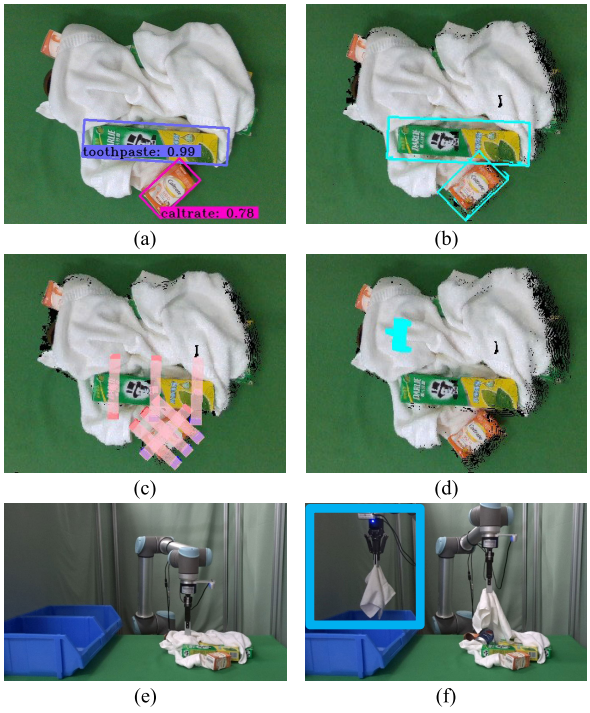



**FIGURE 8.** Entire procedure of picking towel. (a) Detection results. (b) Mapping 2D rotational bounding box to 3D for objects. (c) Collision detection. (d) Obtaining feasible grasping pose. (e) and (f) represent the process of picking a towel.

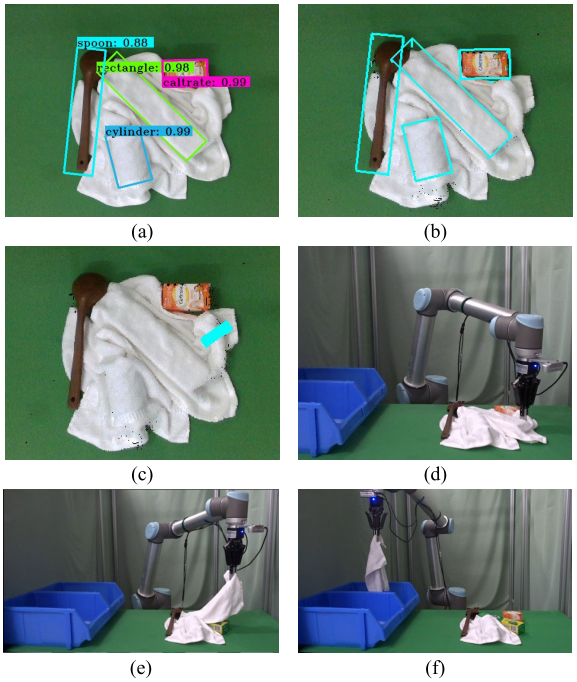



**FIGURE 9.** Entire procedure of picking a towel in certain very special and rare cases, where the geometry of the covered objects is extremely clear. (a) Detection results. (b) Mapping 2D rotational bounding box to 3D for objects. (c) Obtaining feasible grasping pose. (d), (e), and (f) represent the process of picking a towel.

so the point clouds $C\,(rob)$ may not have been consistent with the object. Thus, the additional point clouds could influence the determination of the grasping pose on the towel. Let $C'\,(rob) \subset R^3$ represent the point clouds in $V'\,(rob)$, which replaced $C\,(rob)$ as the point clouds of the rigid objects. Therefore, the practicable grasping region in the point clouds could be expressed as $C\,(g) = C \cap C(sob) \cap C'(rob)$. An overview of grasping a towel is presented in Fig. 9.

## V. CONCLUSIONS

In this study, based on the results of object detection, we introduced a picking strategy that can automatically select a feasible algorithm for grasping pose determination in a mixed situation. Furthermore, it can efficiently solve the grasping of objects mixed with towels. Considering the defects of the traditional $IoU$ and $ArIoU$, we proposed $RIoU$ to measure the accuracy of the detection results. Our method considers $L_{GRIoU}$ as the loss, which can improve the object detection performance. The proposed method can detect objects with different orientations. Using this new detection technique, rigid objects can easily be separated from a mixed scene without segmentation. Furthermore, the grasping pose of the rigid object can be generated directly along the principal axis thereof without using a CAD model or other pose detection methods. The effectiveness of the proposed method was demonstrated in scenes with mixed objects. However, there are some limitations in the current work. One limitation of this work is that the current method does not consider the effect of grasping force. For some soft objects, like bread, if the grasping force is not appropriate, the object will be damaged. The other is that the grasping direction is perpendicular to the table and aligned with a principal axis of the object. However, in practice, there are indeed some situations where it may be more reasonable to grasp objects in other directions. So, in future work, we will study more complex scenes, including other kinds of soft objects, such as bread, bananas, etc. In this case, the grasping force optimization approach will be needed to achieve the balanced grasping of rigid and soft objects. In addition, we will also consider how to realize grasping in more feasible directions to meet more complex scenes.

## REFERENCES

[1] 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.

[2] 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. (CVPR)*, Jun. 2016, pp. 779–788.

[3] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[4] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 21–37.

[6] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*. [Online]. Available: http://arxiv.org/abs/1711.09405

[7] A. Kumar Tanwani, N. Mor, J. Kubiatowicz, J. E. Gonzalez, and K. Goldberg, "A fog robotics approach to deep robot learning: Application to object recognition and grasp planning in surface decluttering," 2019, *arXiv:1903.09589*. [Online]. Available: http://arxiv.org/abs/1903.09589
[8] J. Cai, H. Cheng, Z. Zhang, and J. Su, "MetaGrasp: Data efficient grasping by affordance interpreter network," 2019, *arXiv:1902.06554*. [Online]. Available: http://arxiv.org/abs/1902.06554
[9] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2010, pp. 2308–2315.
[10] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355–3362, Oct. 2018.
[11] X. Wang, X. Jiang, J. Zhao, S. Wang, T. Yang, and Y. Liu, "Picking towels in point clouds," *Sensors*, vol. 19, no. 3, p. 713, Feb. 2019.
[12] Y. Xu, M. Fu, Q. Wang, Y. Wang, K. Chen, G.-S. Xia, and X. Bai, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," 2019, *arXiv:1911.09358*. [Online]. Available: https://arxiv.org/abs/1911.09358
[13] M. Gualtieri, A. ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 598–605.
[14] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *Int. J. Robot. Res.*, vol. 36, nos. 13–14, pp. 1455–1473, 2017.
[15] R. Q. Charles, H. Su, M. Kaichun, 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. 77–85.
[16] H. Liang, X. Ma, S. Li, M. Gorner, S. Tang, B. Fang, F. Sun, and J. Zhang, "PointNetGPD: Detecting grasp configurations from point sets," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3629–3635.
[17] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017, *arXiv:1703.09312*. [Online]. Available: http://arxiv.org/abs/1703.09312
[18] *Cornell Grasping Dataset*. Accessed: Sep. 1, 2013. [Online]. Available: http://pr.cs.cornell.edu/grasping/rect_data/data.php
[19] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1316–1322.
[20] H. Karaoguz and P. Jensfelt, "Object detection approach for robot grasp detection," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 4953–4959.
[21] F. Osawa, H. Seki, and Y. Kamiya, "Unfolding of massive laundry and classification types by dual manipulator," *J. Adv. Comput. Intell. Intell. Informat.*, vol. 11, no. 5, pp. 457–463, Jun. 2007.
[22] C. Bersch, B. Pitzer, and S. Kammel, "Bimanual robotic cloth manipulation for laundry folding," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2011, pp. 1413–1419.
[23] P. Jiménez, "Visual grasp point localization, classification and state recognition in robotic manipulation of cloth: An overview," *Robot. Auto. Syst.*, vol. 92, pp. 107–125, Jun. 2017.
[24] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

**XIAOMAN WANG** received the bachelor's degree in mechanical design manufacture and automation from the Harbin University of Commerce, China, in 2014, and the master's degree in mechatronic engineering from the Harbin Institute of Technology, Shenzhen, China, in 2016, where she is currently pursuing the Ph.D. degree in mechanical engineering. Her research interests include grasping and deformable object manipulation.

**XIN JIANG** (Member, IEEE) received the B.Eng. degree in control engineering from the Dalian University of Technology, Dalian, China, in 2000, and the M.Eng. degree in mechatronics and precision engineering and the Ph.D. degree in aerospace engineering from Tohoku University, Sendai, Japan, in 2004 and 2007, respectively. He was with the Department of Mechanical Engineering, Tohoku University, from 2007 to 2015, as an Assistant Professor. Since 2015, he has been with the Harbin Institute of Technology, Shenzhen, China, where he is currently an Associate Professor with the School of Mechanical Engineering and Automation. His research interests include flexible manipulator, deformable object manipulation, and service robotics.

**JIE ZHAO** received the bachelor's degree in mechanical design manufacture and automation from the Heilongjiang University of Science and Technology, China, in 2015, and the master's degree in mechatronic engineering from the Harbin Institute of Technology, Shenzhen, China, in 2017, where he is currently pursuing the Ph.D. degree in mechanical engineering. His research interests include grasping and in-hand manipulation.

**SHENGFAN WANG** received the B.E.E. degree in automation from Shanghai University, China, in 2018. He is currently pursuing the M.S.E.E. degree in control science and engineering with the Harbin Institute of Technology, Shenzhen, China. His research interests include robotics, deep learning, and computer vision.

**YUN-HUI LIU** (Fellow, IEEE) received the B.Eng. degree from the Beijing Institute of Technology, the M.Eng. degree from Osaka University, and the Ph.D. degree from The University of Tokyo, in 1992. After working at the Electrotechnical Laboratory of Japan as a Research Scientist, he joined The Chinese University of Hong Kong (CUHK), in 1995, where he is currently a Choh-Ming Li Professor of mechanical and automation engineering and the Director of the CUHK T Stone Robotics Institute. He is also an Adjunct Professor with the State Key Laboratory of Robotics Technology and System, Harbin Institute of Technology, China. He has published more than 250 articles in refereed journals and refereed conference proceedings and was listed in the Highly Cited Authors (Engineering) by Thomson Reuters, in 2013. His research interests include visual servoing, medical robotics, multi-fingered grasping, mobile robots, and machine intelligence. He has received numerous research awards from international journals and international conferences in robotics and automation and government agencies. He is the Editor-in-Chief of *Robotics and Biomimetics* and served as an Associate Editor for the IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION and the General Chair for the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems.