applied sciences

MDPI

Article

Deep Learning-Based Occlusion Handling of Overlapped Plants for Robotic Grasping

Mohammad Mohammadzadeh Babr *, Maryam Faghihabdolahi, Danijela Ristić-Durrant and Kai Michels

Institute of Automation, University of Bremen, Otto-Hahn-Allee 1, 28359 Bremen, Germany; m.faghih@iat.uni-bremen.de (M.F.); ristic@iat.uni-bremen.de (D.R.-D.); michels@iat.uni-bremen.de (K.M.)
* Correspondence: babr@iat.uni-bremen.de; Tel.: +49(0)-421-218-62454

Abstract: Instance segmentation of overlapping plants to detect their grasps for possible robotic grasping presents a challenging task due to the need to address the problem of occlusion. We addressed the problem of occlusion using a powerful convolutional neural network for segmenting objects with complex forms and occlusions. The network was trained with a novel dataset named the "occluded plants" dataset, containing real and synthetic images of plant cuttings on flat surfaces with differing degrees of occlusion. The synthetic images were created using the novel framework for synthesizing 2D images by using all plant cutting instances of available real images. In addition to the method for occlusion handling for overlapped plants, we present a novel method for determining the grasps of segmented plant cuttings that is based on conventional image processing. The result of the employed instance segmentation network on our plant dataset shows that it can accurately segment the overlapped plants, and it has a robust performance for different levels of occlusions. The presented plants' grasp detection method achieved 94% on the rectangle metric which had an angular deviation of 30 degrees and an IoU of 0.50. The achieved results show the viability of our approach on plant species with an irregular shape and provide confidence that the presented method can provide a basis for various applications in the food and agricultural industries.

Keywords: deep learning; instance segmentation; occlusion handling; vision-based robotic grasping

Citation: Mohammadzadeh Babr, M.; Faghihabdolahi, M.; Ristić-Durrant, D.; Michels, K. Deep Learning-Based Occlusion Handling of Overlapped Plants for Robotic Grasping. *Appl. Sci.* **2022**, *12*, 3655. https://doi.org/10.3390/app12073655

Academic Editors: Sergiu Dan Stan, Milos Manic and Milos Simonovic

Received: 3 March 2022
Accepted: 1 April 2022
Published: 5 April 2022

1. Introduction

Robotic pick-and-place systems have proven to be essential for increasing production throughput, productivity, and efficiency in numerous industrial applications [1]. The term "robotic pick-and-place" refers to any application in which an object is picked up by a robot at one location, moved, and placed at another location. Traditionally, four main areas of application exist for pick-and-place robots: assembly, packaging, bin-picking, and inspection. In these applications, a target object is grasped by the robot either from a conveyor belt or from a container (bin) and, depending on the application, placed on another conveyor belt, a packaging container, or at another location [2,3]. Such robotic systems are usually equipped with vision systems used for the recognition of target objects to be grasped and moved by the robot. Recent advances in sensor and robotic gripper technology, as well as advances in artificial intelligence (AI), particularly in the area of deep learning (DL), have enabled the use of pick-and-place robots in a broad range of different industrial applications, ranging from traditional applications such as assembly of workpieces to more recent applications in the food industry and agriculture [4–7]. This spread across a wide range of potential applications is characterized by the transition from traditional robotic grasping of exclusively rigid objects, mainly of a standardized size and shape, to the robotic grasping of soft, deformable, and complex-shaped objects. The latter is the case in various applications in agriculture, where complex objects with high variability and heterogeneity, such as fruits and plants, need to be manipulated by the robots. The particular challenge in such robotic applications is dealing with occluded

objects or occluded parts of objects to enable reliable robotic grasping of the target object. Some examples include fruit harvesting [8], plant phenotyping [9], and robotic plant propagation [10]. In the case of fruit harvesting, target crops are often occluded by leaves, and in robotic plant propagation applications where the plant cuttings are transported on conveyor belts, the plants tend to overlap on the conveyor belt. Therefore, vision-based detection of the target objects is a challenging task, which needs to address the problem of occlusion. Once the target object is detected, it is necessary to accurately determine its optimal grasping point so that the robotic grasping can be realized. Depending on the application and vision sensor technology employed, grasping points are calculated directly in a three-dimensional (3D) data-format using the object point cloud [9] or they are first detected in a two-dimensional (2D) single camera image using the result of object segmentation [11]. In the latter case, the 3D coordinates of the object grasping points, as needed for robotic grasping, are calculated using an approach of mapping from a 2D image to 3D working space.

In this paper, novel methods are presented for the detection of target plants among the overlapped plants placed on a flat surface such as a conveyor belt, as well as for finding optimal grasps in a 2D image for possible robotic grasping of the detected target plants. Bearing in mind that considered objects of interest—plants—are flat and flexible (non-rigid) objects with complex shapes, the presented work and the achieved results represent a contribution to further progression in the field of vision-guided robotic grasping of non-rigid objects, which until now has been less researched and developed than robotic grasping of rigid objects [12]. The presented work is divided into two parts. The first presents a novel method for segmenting overlapping plants and classifying the segmented plants into different categories, aiming at the category of target plant cutting that shall be grasped by the robot. The second part presents a novel solution to the problem of finding the optimal robotic grasp of the identified target plant cutting. Solutions to the considered two problems are essential to enable reliable robotic grasping of plants in an application where plants need to be grasped and moved from one position to another, such as robotic plant propagation.

The presented method for target plant detection involves using a high-performance convolutional neural network (CNN) for instance segmentation of objects even in the presence of object occlusion. For training of this CNN, firstly, a dataset from images of plant cuttings placed in random positions on a flat surface with varying degrees of occlusion was created. The dataset is a combination of real and synthetic images, as explained below. Using all available plant cuttings instances from real images, we have developed a framework to synthesize two-dimensional (2D) images as described in Section 3.1. The dataset of overlapping plant cuttings was created using Vaccinium cuttings, which means that the objects of interest are characterized by a thin stem with a number of overlapping leaves, as can be seen in Figure 1.

To grasp the plant cuttings for vision-guided robotic grasping applications, each cutting should first be segmented. For this purpose, the problem of possible occlusions should be addressed. In the work presented in this paper, the occlusion problem is addressed using the method presented in [13], which details a robust network for segmenting objects with complex shapes and severe occlusion. The grasp should be detected for each segmented targeted plant cutting. In the presented system, a single RGB camera is assumed to remain in a fixed position and at a constant distance from the flat surface with the plant cuttings, thus an approach to detect the plant grasp in a 2D RGB image was followed. To cope with the specific shape characteristics of the considered objects of interest, plant cuttings with thin stems and a number of overlapped leaves, we developed a novel effective conventional image processing (CIP)-based method for finding the optimal plant cutting grasp. Figure 1 presents our proposed pipeline for grasp detection of plant cuttings on a flat surface to enable their further grasping by a robot.

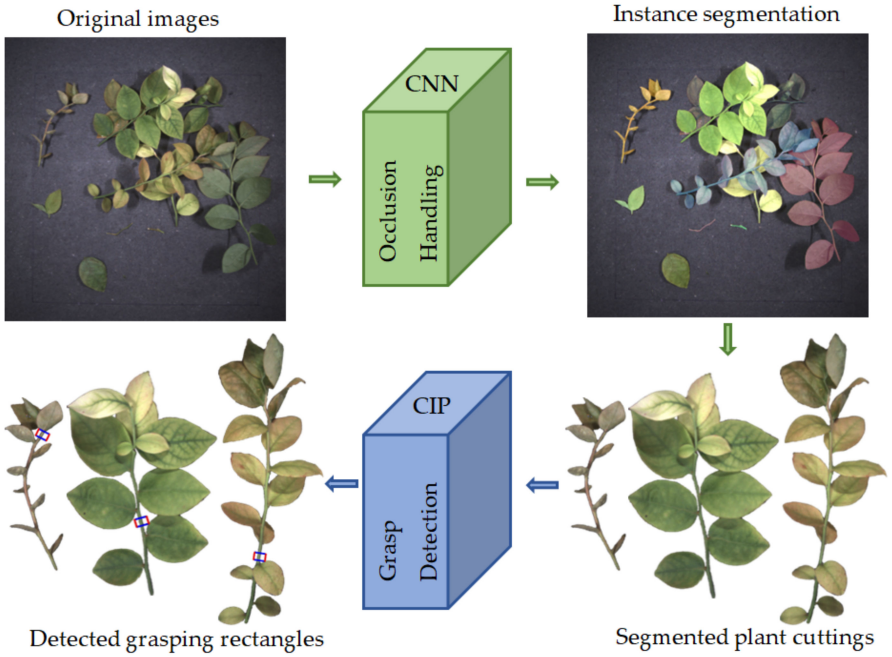


Figure 1. Pipeline of the proposed system for detection of robotic grasps of plant cuttings on a flat surface such as a conveyor belt.

The rest of the paper is organized as follows. In Section 2, an overview of related work is given, with a focus on instance segmentation methods for occlusion handling and detection of target objects' grasps in 2D images based on object segmentation results. Section 3 presents our dataset named the "occluded plants" dataset, as well as the details of the developed methods for segmentation of plant cuttings and detection of grasps. The evaluation results of both methods are given in Section 4.

2. Related Work

Recently, along with many advances in DL, the problem of occlusion in scenes for the instance segmentation tasks has been studied [12–16]. Instance segmentation, a computer vision task for detecting and localizing instances of each object class in an image, represents one of the most challenging problems and encompasses a variety of aspects that are attracting increasing research interest in the computer vision community. Instance segmentation of occluded objects is difficult due to the lack of a complete shape of the object, and only a few methods, such as those presented in [13,14], consider both the complex shape and occlusions of the objects.

The method presented in [17] is a CNN-based method that works with overlapping objects in 2D images and simultaneously segments and classifies objects. Sheared 2D and 3D masks of the overlapping objects are encoded in a volumetric image, and this method performs 3D object segmentation. The main contribution of this network is its ability to generate class-specific instance masks of overlapping biological objects. This method works with 2D images of translucent objects, which makes it easier for the network to lift the label space from 2D to 3D. Presented results in [18] showed that this method achieved high accuracy in overlapping and cluttered situations. However, this method does not suit the solution of the problem of overlapped plants, considered in this paper, as plant objects are not translucent. In [19], a three-layer model is proposed to jointly represent hypotheses, voting elements, instance labels, and their connections for plant imaging analysis. This method deals with partial occlusions as well. The first step in this method is the detection of object centers by Hough voting, and then the instance is segmented around the detected focused point. With updating the Hough votes, all the assignments and weights are updated, and the process is repeated until a stop criterion is met. The possible application of this method for the segmentation of plant cuttings has some limitations. One

of them is the inability to correctly detect plant stems, as the framework of this method is designed for the specific application of plant leaf segmentation. In addition, in cases of heavy occlusion or small leaves, plants' leaves cannot be accurately segmented.

Some existing methods for handling occlusion are dependent on the objects' bounding boxes proposals problem. For example, the method presented in [12] is dependent on the bounding boxes proposal and makes use of image synthesis of objects with occlusion, which is an extension of methods presented in [18,20]. The authors of [12] use a novel relook architecture, which makes use of instance density to segment multi-class masks, extending so the methods already implemented in [21,22]. In addition to segmenting the visible objects, it also generates the invisible parts of the objects. The method presented in [23] is another bounding box-dependent method that performs segmentation and generates the invisible parts of the objects. As this method jointly segments and generates the invisible object's parts, it gets more information about the dependencies of the objects with each other and their occluded regions, their shapes, and appearances. Authors of [24] present a method using oriented boxes instead of axis-aligned boxes for instance segmentation. This method shows that oriented boxes achieve better results and improve the mask predictions especially when the objects are diagonally aligned, overlapping, or touching each other. However, Refs. [12,23,24] considered simple shape objects such as screws, pill bags, and furniture, which are not complex shape objects such as the plant cuttings in our dataset, and therefore the corresponding methods are not suitable for the segmentation problem considered in this paper.

The method given in [14] is a proposal-free method that is capable of dealing with complex shapes with a high number of crossovers. It has shown good performance for small biomedical applications with datasets having complicated shapes and dense crossovers with only one class of objects method rather than a multi-class dataset such as that of our dataset. In addition, this method is computationally expensive, thus the images should be drastically reduced in resolution size, and this means the loss of effective features for segmentation of plant cuttings, which are characterized by a thin stem. Because of these shortcomings, this method is not suitable for the application presented in this paper, segmentation of overlapped plant cuttings placed on flat surfaces.

The network named AdaptIS, presented in [13], is the network that we used in the work presented in this paper, and it is an end-to-end class agnostic method. It takes the raw image as input along with a point on the object and generates a segmentation mask for the object positioned at that point location. It generates pixel-level accurate segmentation and it can also deal with complex object shapes. AdaptIS is not dependent on the object bounding box proposals, and it is performing superior to those that depend on bounding boxes, even better than [24], which makes use of oriented boxes, especially for complex occluded objects. AdaptIS also outperforms "detection first methods" for occluded objects.

Because of cost-effectiveness, real-time performance, and simplicity of vision systems set-up, in a number of vision-based robotic grasping applications, the calculation of object grasps was based on image processing of the 2D images. For example, in [25,26] traditional image processing methods such as contour extraction and morphological operations were used. However, these methods were developed to determine optimal grasps for robotic grasping of fixed structure objects and thus were not suitable for the application considered in this work, which involves robotic grasping of very thin-stemmed plant cuttings that have irregular shapes.

In [11], a combination of DL and conventional image processing was used to detect stem in images and find a grasping point so to enable automatic grasping of the stem and measurement of its diameter of maize and sorghum plants. Since sorghum bears a large and vertical stem, the presented method uses a bounding box detection rather than instance segmentation and then finds the center of the bounding box for grasp point detection. The method does not consider any occlusion and the geometry of the plant is such that there are large empty stem spaces without leaves, so leaf avoidance by the robotic gripper does not need to be considered when determining the grasps. Since the presented

method considers plants with a much simpler complexity than the plants considered in our intended application, it was not possible to apply its straightforward approach for sorghum to our complex plants.

The authors of [27] presented a method for the calculation of grasping points to enable an automatic machine vision-guided grasping system for Phalaenopsis tissue culture plantlets. However, the presented scenario assumes an occlusion-free scene where the segmentation of individual plantlet objects is possible with a simple thresh-old. Starting from a segmented plantlet, a skeleton of the plant is extracted, and the root is distinguished from the leaves, allowing consideration of the middle point of the root as the grasping point. The skeleton of plant cuttings representing the object of interest in our application is much more complex, thus the presented approach could not be followed. This and the above examples illustrate the fact that previously published methods are applicable to specific cases of simple irregular shapes of objects, which cannot be reused for more complex objects, such as complex irregular shape plants in our application. For this reason, a novel method for determining optimal grasps in 2D images had to be developed, as explained in Section 3.4.

3. Materials and Methods

3.1. Dataset

The dataset of overlapping plant cuttings placed on a flat surface such as a conveyor belt was created using Vaccinium cuttings. Two unique appearances of each cutting, front and back, were used to create different overlap cases of the cuttings on the flat background. The RGB images of the overlapped cuttings with a resolution of 2048 $\times$ 1536 pixels were captured with a stereo camera set. In total, 650 images were captured, and they were annotated for instance segmentation.

Since plants do not have a fixed structure, their overlap can be so complicated that in some cases it is not possible even for human annotators to distinguish between different plant instances. In the context of the present work, several terms are defined to narrow down the cases of occlusion:

- An occlusion patch is a group of pixels in the image for which one part of a plant cutting is occluded by other plant cuttings and, thus, this part is not visible in the image. Based on this definition, one cutting can have several occlusion patches with a single cutting or multiple other cuttings.
- A normal occlusion patch is the occlusion patch in the image for which a part of only one plant cutting is not visible. It should be noted that one plant cutting can have multiple normal occlusion patches. The red windows in Figure 2 illustrate some locations where normal occlusion patches are present.
- A complex occlusion patch is the occlusion patch in the image for which parts of more than one plant cuttings are not visible. One plant cutting might have multiple complex occlusion patches or even a mixture of several normal and complex occlusion patches. The purple windows in Figure 2 illustrate some locations where complex occlusion patches are present.
- A normal occlusion image is an image that contains at least one normal occlusion patch and no complex occlusion patches.
- A complex occlusion image is an image that contains at least one complex occlusion patch.

Figure 3 shows examples of normal and complex occlusion images with annotations of plant cuttings as belonging to different object classes. As it can be seen, the following four object classes are introduced: Target Cutting, Occluded Cutting, Singularized Cutting, and Remains. The cuttings that occlude other cuttings and that are not occluded themselves as they are on the top of other cuttings are classified as Target Cuttings. The cuttings which are occluded by other cuttings are classified as Occluded Cuttings. The cuttings that are not occluded and also do not occlude other cuttings are classified as Singularized Cuttings. Individual plant parts such as a single leaf or a stem part or cuttings that have two or fewer

leaves are classified as Remains. In total, 25% of all images in the presented dataset are complex occlusion images.

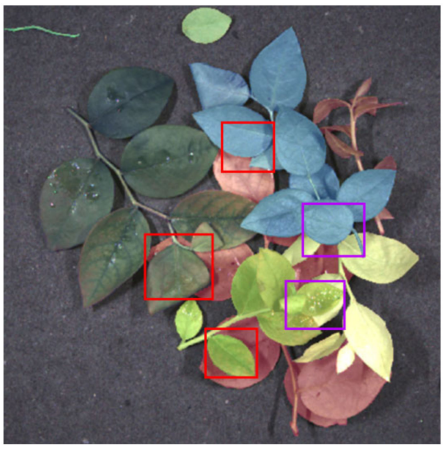


Figure 2. Example of an image from the occluded plants dataset with normal and complex occlusion patches. The red windows illustrate some locations where normal occlusion patches are present, and the purple windows show some locations of complex occlusion patches.

(**a**) (**b**)

Figure 3. Examples of annotated images from the occluded plants dataset. (**a**) Normal occlusion image. (**b**) Complex occlusion image.

The overlap of sum (OoS) [28] metric can indicate the occlusion and crowding level of images in a dataset. This metric OoS for an image of a dataset is defined as follows:

$$\text{OoS} = \begin{cases} 1 - \frac{\left|\cup_{i=1}^{n} C_i\right|}{\sum_{i=1}^{n} |C_i|}, & n > 0 \\ 0, & n = 0 \end{cases} \quad (1)$$

where C_i is the area of the bounding box (bbox) or convex hull (convex) of instance i in an image, n is the number of instances in an image, and $\cup$ is the union operation [28]. Table 1 presents the average OoS over all images of the dataset of occluded plants presented in this paper, as well as for several standard datasets.

Table 1. Comparison of the OoS measure with different datasets containing images with occlusions.

Dataset	Bbox	Convex
COCO [28,29]	0.14	0.07
Cityscapes [28,30]	0.15	0.09
OC Human [15,28]	0.25	020
Occluded plants (normal occlusion)	0.19	0.13
Occluded plants (complex occlusion)	0.28	0.20

As the OoS metric in Table 1 shows, images of the presented occluded plants dataset contain by far more occlusions than the COCO and Cityscape datasets. In the case of complex occlusions, the occluded plants dataset has an even higher degree of occlusion than the OC Human dataset, which focuses on heavily occluded people in crowded scenes.

3.2. Synthesizing 2D Images

A novel framework for synthesizing 2D images was developed using the ground truth of real RGB images. In this approach, the instances of the plant cuttings derived from the annotation of real RGB images were firstly transformed by applying some geometric transformations such as rotation and scaling, after which they were placed in a random order on empty background images (i.e., images of flat surfaces without any objects placed on them). The plant cutting instances belonging to the Singularized Cutting and Target Cutting classes were used for generating synthetic images, since the instances of these classes have a complete shape of the cuttings in existing real RGB images. The algorithm for image synthesis was designed in such a way that the entire process did not follow any particular pattern. This means that the generated synthetic images do not have any similarities in the overlapping cases. In this approach, the ground truth annotations of synthetic images were performed in parallel with the process of creation of synthetic images themselves. The framework generates images with both normal and complex occlusions, and the percentage of complex occlusion images can be set in advance. 5000 synthetic images were generated, including 30% complex occlusion images. Figure 4 shows an image resulted from the process of synthesizing 2D images (the left bottom image in Figure 4) and several intermediate views of the synthesis process. The intermediate views correspond to intermediate steps of addition of the plant cuttings to the viewed scene. The example resulted synthetic image is a complex occlusion image.

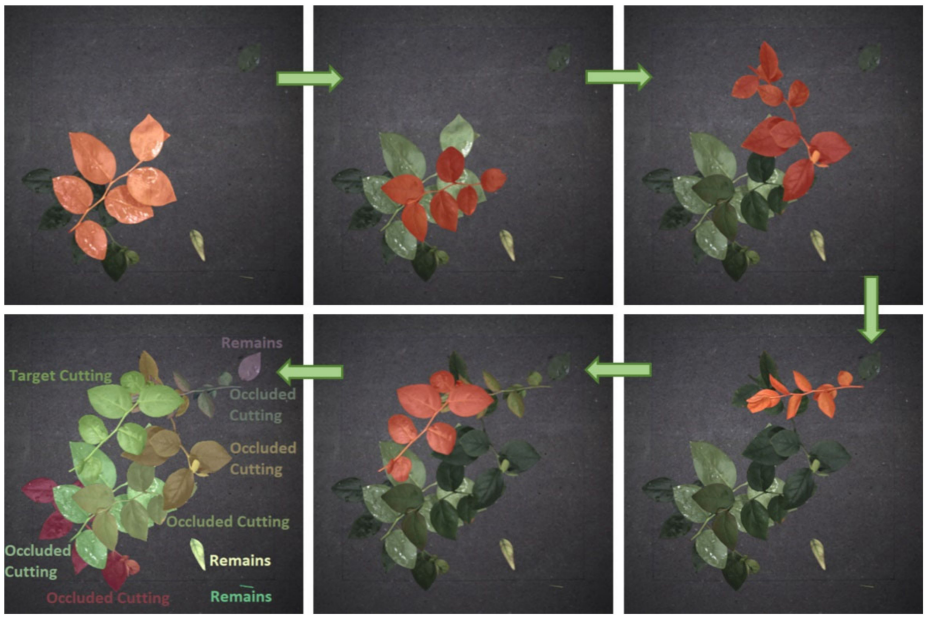


Figure 4. A synthesized image and several intermediate views of the synthesis process.

3.3. Occlusion Handling

AdaptIS is essentially a class-agnostic instance segmentation method that can perform multi-class instance segmentation or a panoptic segmentation [31] using a standard semantic segmentation pipeline [13]. Since the occluded plants dataset is a multi-class dataset, we implemented AdaptIS to perform panoptic segmentation that unifies semantic and instance segmentation, thereby assigning a class label to each pixel and recognizing and segmenting each object instance. AdaptIS uses point proposal to generate masks of objects located at the proposed points. It creates an object mask for each proposed point without any heuristic post-processing [13]. Figure 5 shows the instance segmentation masks produced by AdaptIS for different proposed points on the plant cutting objects in the top right image. As it can be seen, for the points corresponding to the same object, AdaptIS creates very similar masks.

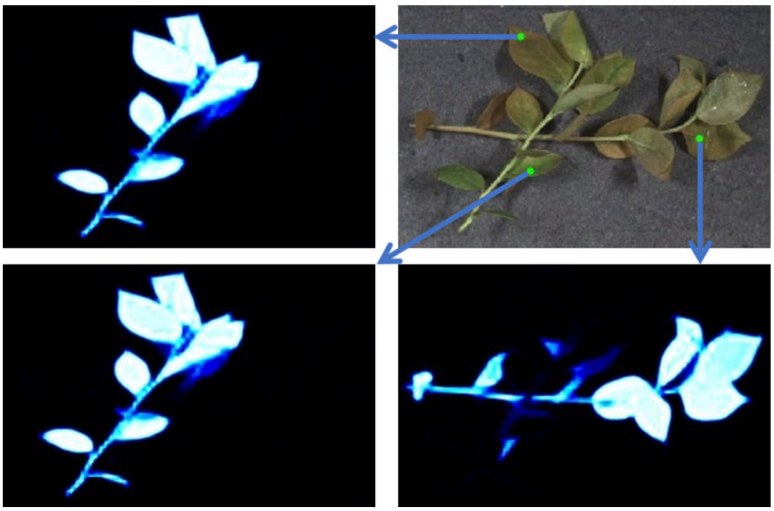


Figure 5. Instance segmentation masks produced by AdaptIS for various point proposals.

AdaptIS optimizes a target loss function for a given image I and a proposed point at image coordinate (x, y). It uses a pixel-wise loss function for comparing the prediction with the mask of the target object located at the point (x, y). Since AdaptIs provides pixel-precise segmentation, it is suitable for objects with complex shapes and can handle cases of heavy occlusion [13].

AdaptIS includes a specific head for point proposal that proposes about 100 points per image in the inference time. The instance segmentation output of AdaptIS at inference time is a pixel-wise mask for a single object. For segmenting all objects in the image, different point proposals are made to create masks for multiple objects in sequence. The iteration only needs to be conducted for a lightweight AdaptIS head, while the backbone should be run once. Moreover, after creating a mask for a proposed point, all point proposals located in the created mask are excluded from the set of point proposals. These two techniques significantly reduce the computation time and make the iterative method applicable in practice [13].

3.4. Grasp Detection

We developed a novel method for detecting the robotic grasp of a segmented plant cutting based on Conventional Image Processing (CIP). In general, a grasp specifies how a robot end-effector (gripper) can be arranged to safely grip an object and lift it without slipping it off. A grasp holds the information about the grasping point, the grasping orientation, and the opening width of the robot gripper [32]. Since the objective of the intended robotic grasping application is to grasp the plant cuttings classified as Target Cuttings and Singularized Cuttings (see Section 3.1), the developed method was designed to detect the grasps for full plant cuttings and not for the Occluded Cuttings or Remains.

Since the plant cuttings are small and delicate, a small two-finger gripper or tweezers with adequate distance between the tips can serve as a gripper for grasping the plant cuttings from flat surfaces. The oriented rectangle [33] is used here for the grasp representation, which indicates the position and orientation of a two-finger gripper before closing on an object:

$$g = \{x, y, \theta, h, w\} \quad (2)$$

where (x, y) is the center of the rectangle in the image coordinate system, θ is the orientation of the rectangle with respect to the horizontal axis of the image coordinate system, h is the height of the rectangle or, here, it is the thickness of each finger that is a constant value, and w is the width or the distance between the two robotic gripper fingers. The grasp representation with a pair of points [34], where each point is the end position of each finger of the gripper, was also used during the grasp detection process to find the optimal oriented rectangle. To succeed in an actual grasp of a plant cutting, the predicted grasps must meet the following criteria:

- The center of the gripper (center of the oriented rectangle) must be on the stem of the plant cutting.
- The orientation of the predicted grasp should be aligned with the direction of the plant cutting's stem.
- The grasping point must have sufficient distance to the plant leaves to accommodate the open gripper and to avoid collision with leaves.

The pipeline of the developed algorithm for grasp detection is shown in Figure 6 and the algorithm is explained in detail in the following subsections.

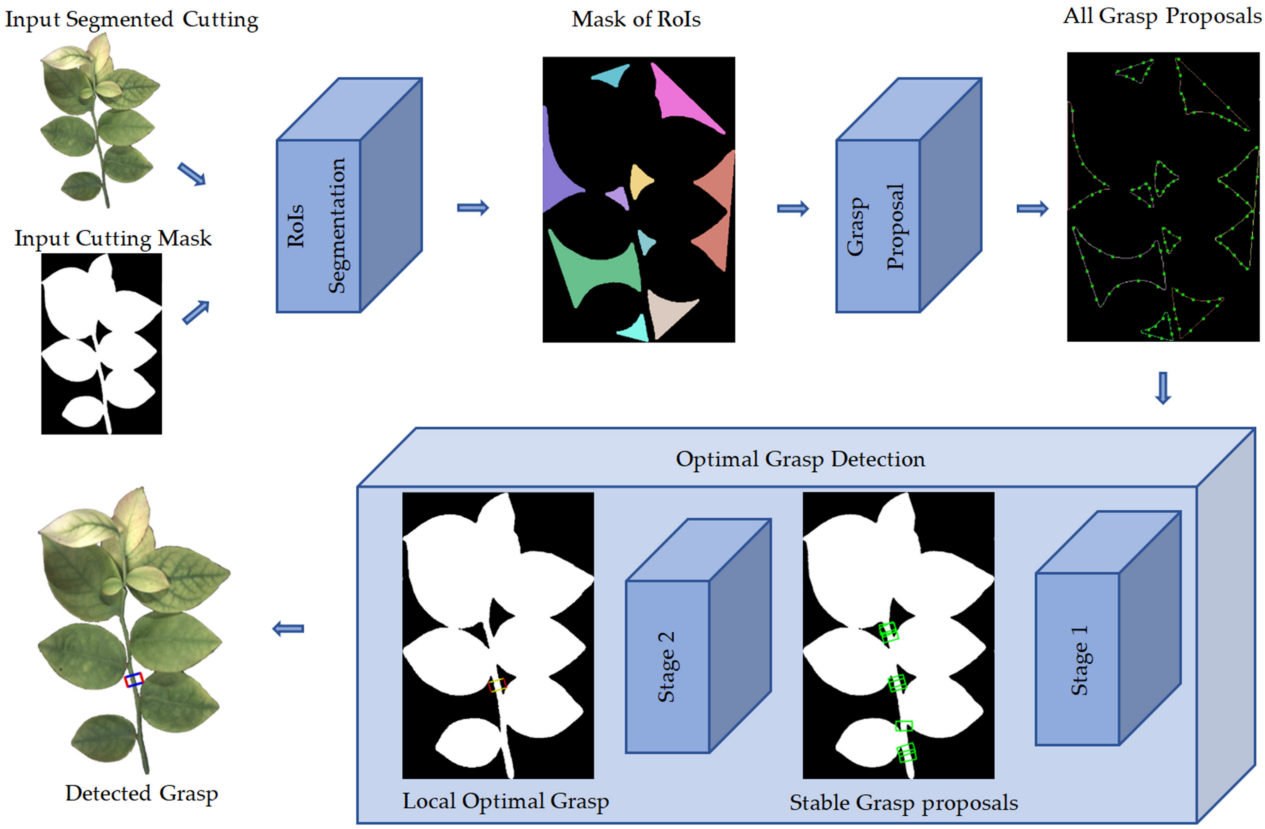
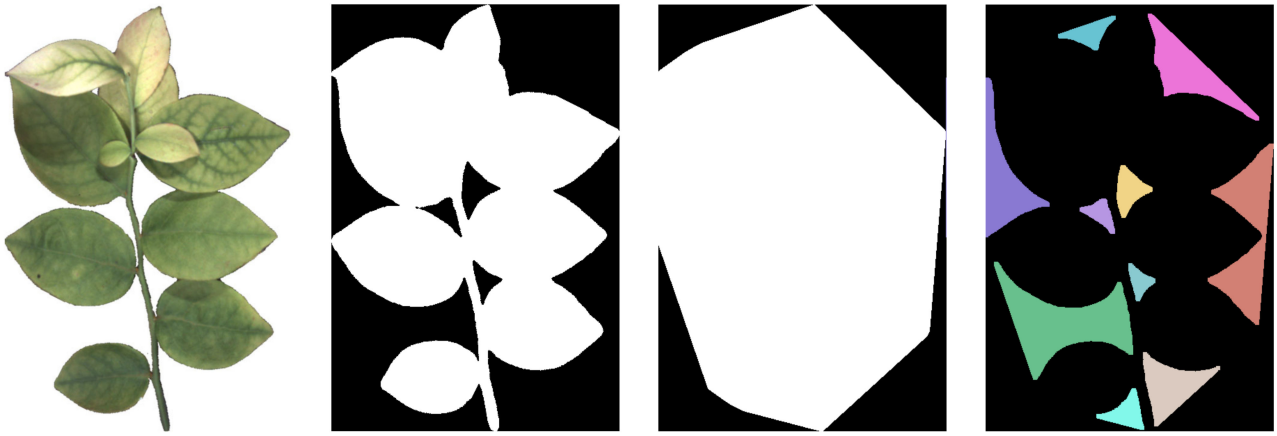


Figure 6. Pipeline of the proposed algorithm for detection of the robotic grasp of a segmented plant cutting.

3.4.1. Regions of Interest (RoIs) Segmentation

RoIs are regions in the bounding box of a plant cutting where no plant parts are present, and robotic gripper fingers should be placed in two different RoIs to grasp the cutting. The method employed to determine RoIs is based on the method presented in [26]. The input to the developed algorithm is the binary segmentation mask of a full plant cutting. Firstly, at preprocessing stage, multiple morphological transformations and filters (Opening, Dilation, and Smoothing) are applied to the segmentation mask to smooth its boundaries and to fill small gaps in the mask. The plant cuttings inherent feature that their stems usually are lying along the longest side of the cutting's bounding box was used for the subsequent steps by rotating the segmentation mask so that its longer side becomes vertical (Figure 7b).



(**a**) (**b**) (**c**) (**d**)

Figure 7. RoIs segmentation pipeline. (**a**) Segmented plant cutting extracted from an original RGB image. (**b**) Segmented plant cutting's mask after preprocessing. (**c**) Convex hull mask. (**d**) RoIs mask where each color represents one RoI.

In order to identify the RoIs for a plant cutting mask, first, the convex hull of the cutting mask is determined, and a binary mask is generated for it. Afterward, the plant cutting mask (Figure 7b) is subtracted from the convex hull mask (Figure 7c), and an opening morphological transformation is applied to the mask, resulted from performed subtraction, to separate the individual RoIs. By finding contours on the resulted mask, all the RoIs can be segmented separately. Figure 7 illustrates the RoIs segmentation pipeline.

3.4.2. Grasp Proposal

Representation of the grasp by a pair of points is used to determine the grasp proposals. As mentioned in the previous section, the robotic gripper fingers should be placed in two different RoIs to grasp the plant cutting, which means that each point of a grasp should also be in separate RoIs. The stem of a plant cutting is the suitable part for possible stable robotic grasping of the cutting because it is firm enough to withstand the weight of the plant cutting after grasping in addition to the pressure of the robotic gripper fingers. For plant cuttings, the edge of RoIs is suitable for the choice of a point from the grasp pair of the points because it offers features to determine whether the grasp is on the stem of the cutting or not. We consider several points on the contour of each RoI, and in principle, any pair of points from two separate RoIs represent a proposed grasp. However, only a few of the proposed grasps can succeed in actual plant cutting grasping.

3.4.3. Optimal Grasp Detection

To find the optimal grasp among the grasp proposals, in the first stage, the unstable grasps are filtered out, and in the second stage the orientation of the remaining grasp

proposals is refined, then the grasps are evaluated against a defined optimality measure to find the optimal grasp.

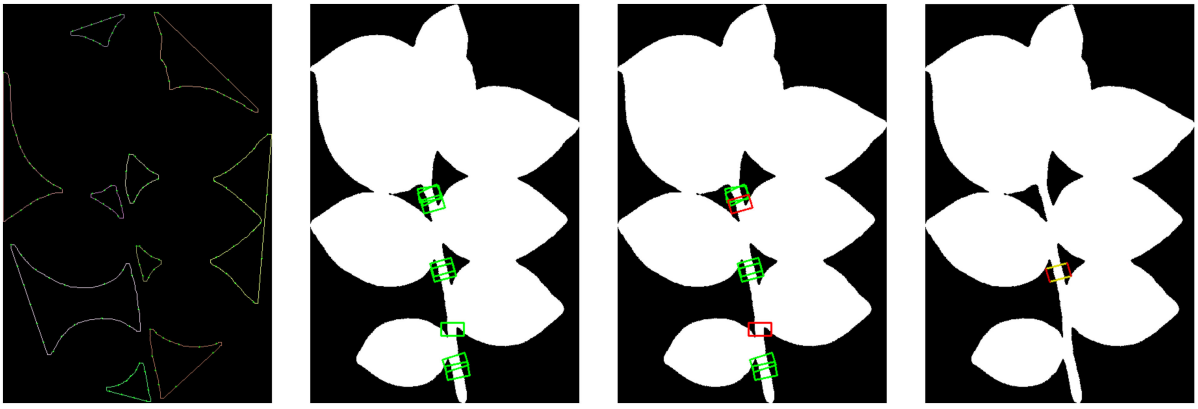
To filter out the most likely unstable grasps, the following two characteristics of plant cuttings are considered. First, the distance between the two points of each proposed grasp should be within an acceptable range determined by the minimum and maximum possible width of the cutting's stem in the pixel unit. Second, the so-called "Cutting's Interested Region" (CIR) is defined as a region in the bounding box of a plant cutting where all possible grasps should be placed in. Furthermore, CIR is a portion of the bounding box of the plant cutting, and its center is aligned with the center of the bounding box. Since the plant cuttings are oriented vertically and the stems are almost in the middle of the bounding boxes of the cuttings, the horizontal ratio of the CIR and the bounding box of the cuttings should be much smaller than the vertical one. The set ratios between the CIR and the bounding box of the cuttings for horizontal and vertical sides are 0.5 and 0.9, respectively.

After applying the above constraints to the grasp proposals, the bulk of the unstable grasps are filtered out. In the next step, the orientation of the remaining grasp proposals is refined to be aligned with the direction of the plant cutting's stem. Since the plant cutting's stem is roughly cylindrical, the orientation of a grasp proposal would be aligned along the stem if the distance between the points of the grasp is approximately as large as the width of the stem. To refine the orientation of a grasp proposal, one point of the grasp remains fixed and another is moved along the stem to minimize a function that calculates the distance between two points of a grasp proposal. The remaining grasp proposals are refined based on this method to increase the precision of the grasps' orientation.

Finally, for finding the optimal grasp, we defined a metric called Overlapped Grasp Index (OGI). This metric uses both the pair of points and the rectangle representations of grasps.

$$\mathrm{OGI} = \frac{Grasp\ rectangle \cap Plant\ Cutting\ mask}{(Distance\ of\ the\ points\ of\ the\ grasp) \times (Grasp\ height)} \tag{3}$$

OGI indicates the overlap of the grasp rectangle with the leaves of the cutting. The minimum, indeed the optimal value of this metric, is 1, and the higher values show the overlap of the grasp with the cutting's leaves. Thus, the proposed grasp with the smallest OGI is the optimal local grasp for the cutting. Figure 8 shows the key intermediate results of the grasp proposal and optimal grasp detection pipeline.



(**a**) (**b**) (**c**) (**d**)

Figure 8. Key intermediate results of the grasp proposal and optimal grasp detection pipeline. (**a**) All grasp proposals are marked as green points on the boundaries of ROIs. In principle, any pair of the points from two separate RoIs is a proposed grasp (**b**) Grasp proposal after filtering out the most likely unstable grasps. (**c**) Refined proposed grasps. Grasps marked with red rectangles are excluded due to overlap with leaves as determined by high OGI. (**d**) Local optimal grasp, which has the smallest OGI.

4. Results and Discussion

4.1. Occlusion Handling

The occluded plant dataset was split into sets of 500, 50, and 100 real images for training, validation, and testing, respectively. The test dataset itself included 50 normal occlusion images and 50 complex occlusion images. In addition to the real images, we included 5000 synthetic images in the training set. AdaptIS was trained through a two-stage training procedure. The backbone and segmentation head were first trained for 230 epochs, then these parts were frozen, and only the point proposal head was trained for additional 15 epochs. The first stage consisted of 220 epochs of training with real and synthetic images, followed by 10 epochs of fine-tuning training with real images. In the second stage of the training procedure, only the real images were used.

For the backbone, we used ResNet50 [35] and trained the network using Adaptive Moment Estimation (Adam) Optimizer with 2 GPUs and batch size 4. At both training stages, the base learning rate was 0.0005 and it was reduced by the cosine learning rate scheduler. Each original image contained a region of interest with a resolution of 1536 × 1536 pixels, such that all cutting instances were always placed in this region of interest. The input size of the network was 512 pixels, thus, downscaling the original image from 1536 to 512 pixels resulted in a drastic loss of information. To reduce this downscaling in the training step, first, the part of the image where the plants were placed was cropped out and then scaled down to 512 pixels. An additional advantage of this method is that it acts as a random scaling augmentation in the range of (0.33 to 0.66). Apart from scale augmentation, we also used random rotation, brightness, contrast, and color data augmentation.

Table 2 presents the panoptic segmentation result of AdaptIS for normal and complex occlusion. The standard metrics of panoptic segmentation are Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ) [31]. Each of the panoptic metrics is split into metrics for the Stuff (PQ^{St}, SQ^{St}, RQ^{St}) and Thing classes (PQ^{Th}, SQ^{Th}, RQ^{Th}) [31]. In the presented work, only the Image Background represented a Stuff class, and the results for that were about 100 on all metrics, so we present here the panoptic segmentation result only for Thing classes to show more clearly the performance of AdaptIS on the occluded plants dataset. The optimal input size of the network in inference was 640 pixels. The inference time was 650ms per image on the used local machine, with a single RTX 2080 Ti GPU, MXNET 1.7, and CUDA 10.2.

Table 2. Panoptic segmentation results with AdaptIS on the presented occluded plants dataset.

Classes	Normal Occlusion			Complex Occlusion		
	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{Th}	SQ^{Th}	RQ^{Th}
Singularized Cutting	85.8	93.6	91.7	89.1	94.5	94.4
Remains	91.0	94.0	96.9	86.0	94.1	91.4
Occluded Cutting	76.5	88.5	86.5	72.0	83.2	86.5
Target Cutting	78.2	89.9	87.1	86.8	91.4	95.0
Average	82.9	91.5	90.6	83.5	90.8	91.8

Surprisingly, as the results in Table 2 show, the accuracy of AdaptIS for Occluded and Target Cutting classes is very close for the normal and complex occlusion, confirming the robustness of AdaptIS in segmenting severe occluded scenes. As expected, the Occluded Cutting instances are the most challenging class for segmentation, and the accuracy of the network is lower for them than for other classes. Figure 9 shows examples of panoptic segmentation results with AdaptIS for normal and complex occlusion. It can be seen that the Occluded Cutting instances in almost all examples are split into multiple parts, or the large portion of a cutting in the example at the bottom left is hidden. Although these are difficult cases for a segmentation task, AdaptIS successfully handled them and accurately segmented these Occluded Cuttings. The network moreover makes a precise distinction between the Target cutting and Occluded cuttings in the cases of complex occlusions.

(**a**) (**b**) (**c**) (**d**)

Figure 9. Examples of panoptic segmentation results with AdaptIS network on occluded plants dataset. (**a**) Ground truth of normal occlusion images. (**b**) Segmentation results of normal occlusion images. (**c**) Ground truth of complex occlusion images. (**d**) Segmentation results of complex occlusion images. In order to have a clearer visualization of cutting masks, the mask of the Stuff classes (here the image background) is not visualized.

Generally, the stems of the plant cuttings are thin, thus it is a challenge for the segmentation networks to segment them completely and without any discontinuities. This makes the stems of the cuttings the most difficult part of the plant cuttings for segmentation. Since the input size of the AdaptIS network is 640 pixels, the width of the stems ranges from 2 to 7 pixels, which makes their segmentation even more difficult. As the sample results in Figure 9 show, the stems of the plant cuttings are segmented by AdaptIS completely and without any discontinuities, which proves the robustness of this network in segmentation.

The goal of the intended robotic grasping application is to grasp the instances of Target Cuttings and Singularized Cuttings. Therefore, the segmentation masks of plants of these classes are used for grasp detection, and the precision of these masks impacts the result of grasp detection. As Table 2 shows, the SQ, i.e., the percentage of average IoU (Intersection over Union) of the matched segments, for these two classes is around 90%, which is a high precision for the purpose of our intended robotic grasping. However, in some rare cases such as in the top left example in Figure 9, the segmentation result of a Target Cutting instance is incorrect, and as it can be seen, one leaf is inaccurately segmented, which may affect the accuracy of the grasp detection for this instance.

4.2. Grasp Detection

The presented grasp detection method was applied to 100 images of the test dataset, which included around 250 instances of grasping target classes (Target Cutting and Singularized Cutting). Some features used for grasp detection are related to the resolution of

the image. Two of these features are the minimum and maximum possible width of the plant cutting stem and the minimum acceptable area of RoI. Since the optimal input size of the AdaptIS for inference was 640 pixels, these features were set for an image resolution of 640 pixels. The range of the cutting stem was set to 2 to 7 pixels, and the smallest acceptable area of the RoI was set to 125 pixels. As mentioned in Section 3.4.3, the CIR value, which is independent of the image resolution, was set to 0.5 and 0.9 for the horizontal and vertical sides of the CIR, respectively.

To evaluate the achieved results in grasp detection, we used the rectangle metric [36]. Based on this metric, a predicted grasp is correct if the deviation of grasp orientation between the predicted grasp and the ground truth grasp is less than 30° and the IoU (intersection over union) value between the predicted rectangle and the ground truth is more than 0.5. To analyze the performance of our grasp detection method, we used the rectangle metric at different orientation angle criteria (5, 15, 30, and 45 degrees) and different IoU criteria (0.25, 0.5, and 0.75). The runtime was 15ms per image on the local machine with an Intel® Xeon® Silver 4216 processor. Table 3 presents the evaluation results of grasp detection on the test dataset.

Table 3. Evaluation results of grasp detection on the test dataset.

IoU	Rectangle Metric [%]			
	Angle 5°	**Angle 15°**	**Angle 30°**	**Angle 45°**
0.25	68	93	95	95
0.50	67	92	94	94
0.75	50	66	67	67

As the result in Table 3 shows, the standard rectangle metric for the presented method is 94%, which is a high performance and 50% of the predicted grasps are correct in the most constrained case, namely with an angular deviation of 5° and an IoU of 0.75, demonstrating the high robustness of our method for grasp detection of plant cuttings. Conversely, the predicted grasps that belong to the least restricted case (angular deviation 45° and IoU 0.25) lead to successful grasping but also cause rotation of the plant cuttings due to the moderate angular deviation, and the accuracy of our method in this regard is 95%.

As the examples in Figure 10 show, the predicted grasps are on the stem with sufficient clearance to accommodate the open gripper as well as with a small angular deviation. In Figure 10a, the segmented Target Cutting at the bottom left corner has a partial segmentation inaccuracy. Despite this inaccuracy, the grasp is correctly predicted, demonstrating the robustness of our method for grasp detection.

The stem around the tip of the plant cutting is thinner and softer, thus, grasping and picking the cutting at this point may cause bending the cutting toward the ground, though it will not lead to falling. In Figure 10b, the grasp shown in the bottom left corner is an example of this rare occurrence. The reason that such cases are rare is that the tips of cuttings usually have several dense leaves, thus the stem around the tip of the cutting does not have enough room to accommodate the gripper. For this reason, in the last step of our proposed method, most of the grasp proposals near the tip of the cutting are excluded.

(**a**) (**b**)

Figure 10. Examples of grasp detection on the test dataset. (**a**) Normal occlusion image. (**b**) Complex occlusion image.

5. Conclusions

In this paper, we presented a novel method for determining grasps for robotic grasping of plant cuttings from a flat surface such as a conveyor belt. The method consists of using a CNN for handling plant occlusions when segmenting individual plant cuttings and detection of optimal grasps based on the segmentation results of the CNN. We compared the performance of the used CNN, AdaptIS, using the custom-made occluded plants dataset with two different levels of occlusion severity. The results show that AdaptIS can accurately segment the overlapping plants and its performance is robust at various levels of occlusion. Moreover, although precise segmentation of the stem of a plant cutting represents a challenging task for the segmentation networks, the achieved segmentation result of AdaptIS for this part of the plant cuttings was complete and free from any discontinuities. AdaptIS segmented the target classes for the robotic grasping (Target Cuttings and Singularized Cuttings) with 90% accuracy according to the SQ metric, which is promising for the purpose of our intended robotic grasping. Our CIP-based method for grasp detection of plants achieved 94% for the rectangular metric, and in fact, 50% of the predictions had the highest accuracy in the evaluation range, i.e., an angular deviation of 5 degrees and an IoU of 0.75. We have shown the feasibility of our approach on a plant genus with an irregular object shape similar to many other plant genera in nature so that our method could build a basis for different applications in the food and agricultural industries.

Our future work will focus on the implementation of the presented method in a robotic pick-and-place system and conducting field tests of the system, with the objective of picking the identified target plant cuttings from the pile of plant cuttings to enable their sorting and placing them to another station of a robotic plant propagation system. In this way, future evaluation of the proposed method will be performed by evaluating the success rate of robotic grasping. In addition to the implementation and subsequent evaluation of the presented method in practice, our future work will also involve its further improvement. Namely, as the occlusion is a 3D feature, using the depth information can improve the

occlusion handling accuracy, thus the occlusion handling will be investigated with the segmentation networks using RGB-D images. Moreover, our future work will focus on the generalization of the presented method by extending the training dataset with images of different plant genera, so that the limitation of applicability of the presented methods to only one considered plant genus is overcome.

Author Contributions: Conceptualization, M.M.B. and M.F.; methodology, M.M.B. and M.F.; software, M.M.B.; validation, M.M.B.; formal analysis, M.M.B. and M.F.; investigation, M.M.B.; resources, M.M.B.; data curation, M.M.B.; writing—original draft preparation, M.M.B., M.F.; writing—review and editing, M.M.B., M.F. and D.R.-D.; visualization, M.M.B. and M.F.; supervision, M.M.B., D.R.-D. and K.M.; project administration, M.M.B., D.R.-D. and K.M.; funding acquisition, D.R.-D. and K.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work is part of the research project funded by the Bremer Aufbau-Bank (BAB) and the EFRE (European Funds for Regional Developments) EU Investments in Future of Bremen under contract VE0118B.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The custom dataset for our proposed method due to confidentiality cannot be publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Javaid, M.; Haleem, A.; Singh, R.P.; Suman, R. Substantial capabilities of robotics in enhancing industry 4.0 implementation. *Cogn. Robot.* **2021**, *1*, 58–75. [CrossRef]
2. Fujita, M.; Domae, Y.; Noda, A.; Ricardez, G.A.G.; Nagatani, T.; Zeng, A.; Song, S.; Rodriguez, A.; Causo, A.; Chen, I.M.; et al. What are the important technologies for bin picking? Technology analysis of robots in competitions based on a set of performance metrics. *Adv. Robot.* **2020**, *34*, 560–574. [CrossRef]
3. Han, S.D.; Feng, S.W.; Yu, J. Toward Fast and Optimal Robotic Pick-And-Place on a Moving Conveyor. *IEEE Robot. Autom. Lett.* **2020**, *5*, 446–453. [CrossRef]
4. Arents, J.; Greitans, M. Smart Industrial Robot Control Trends, Challenges and Opportunities within Manufacturing. *Appl. Sci.* **2022**, *12*, 937. [CrossRef]
5. Bader, F.; Rahimifard, S. A methodology for the selection of industrial robots in food handling. *Innov. Food Sci. Emerg. Technol.* **2020**, *64*, 102379. [CrossRef]
6. Atefi, A.; Ge, Y.; Pitla, S.; Schnable, J. Robotic Technologies for High-Throughput Plant Phenotyping: Contemporary Reviews and Future Perspectives. *Front. Plant Sci.* **2021**, *12*, 1082. [CrossRef]
7. Bac, C.W.; Hemming, J.; van Henten, E.J. Stem localization of sweet-pepper plants using the support wire as a visual cue. *Comput. Electron. Agric.* **2014**, *105*, 111–120. [CrossRef]
8. Jiao, Y.; Luo, R.; Li, Q.; Deng, X.; Yin, X.; Ruan, C.; Jia, W. Detection and localization of overlapped fruits application in an apple harvesting robot. *Electronics* **2020**, *9*, 1023. [CrossRef]
9. Joffe, B.; Ahlin, K.; Hu, A.-P.; Mcmurray, G. Vision-guided robotic leaf picking. *EasyChair Prepr.* **2018**, *250*, 1–6. [CrossRef]
10. Integrating Computer Vision into Horticulture Robots—Robovision. Available online: https://robovision.ai/case-study/iso-group-case-study/ (accessed on 22 February 2022).
11. Atefi, A.; Ge, Y.; Pitla, S.; Schnable, J. Robotic Detection and Grasp of Maize and Sorghum: Stem Measurement with Contact. *Robotics* **2020**, *9*, 58. [CrossRef]
12. Wada, K.; Kitagawa, S.; Okada, K.; Inaba, M. Instance Segmentation of Visible and Occluded Regions for Finding and Picking Target from a Pile of Objects. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Madrid, Spain, 1–5 October 2018; pp. 2048–2055. [CrossRef]
13. Sofiiuk, K.; Barinova, O.; Konushin, A. AdaptIS: Adaptive Instance Selection Network. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop, Seoul, Korea, 27–28 October 2019; pp. 7354–7362. [CrossRef]
14. Hirsch, P.; Mais, L.; Kainmueller, D. PatchPerPix for Instance Segmentation. In *Computer Vision—ECCV 2020, Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23—28 August 2020*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12370, pp. 288–304. [CrossRef]
15. Zhang, S.-H.; Li, R.; Dong, X.; Rosin, P.; Cai, Z.; Han, X.; Yang, D.; Huang, H.; Hu, S.-M. Pose2Seg: Detection Free Human Instance Segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 889–898. [CrossRef]

16. Salvador, A.; Bellver, M.; Campos, V.; Baradad, M.; Marques, F.; Torres, J.; Giro-i-Nieto, X. Recurrent Neural Networks for Semantic Instance Segmentation. *arXiv* **2017**, arXiv:1712.00617.
17. Böhm, A.; Ücker, A.; Jäger, T.; Ronneberger, O.; Falk, T. ISOODL: Instance segmentation of overlapping biological objects using deep learning. In Proceedings of the International Symposium on Biomedical Imaging, Wasington, DC, USA, 4–7 April 2018; pp. 1225–1229. [CrossRef]
18. Georgakis, G.; Mousavian, A.; Berg, A.C.; Košecká, J. Synthesizing training data for object detection in indoor scenes. *Robot. Sci. Syst.* **2017**, *13*. [CrossRef]
19. Yu, J.-G.; Li, Y.; Gao, C.; Gaoa, H.; Xia, G.-S.; Yub, Z.L.; Lic, Y.; Gao, H.; Yu, Z.L.; Li, Y. Exemplar-Based Recursive Instance Segmentation with Application to Plant Image Analysis. *IEEE Trans. Image Process.* **2020**, *29*, 389–404. [CrossRef]
20. Dwibedi, D.; Misra, I.; Hebert, M. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1310–1319. [CrossRef]
21. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 386–397. [CrossRef]
22. Do, T.T.; Nguyen, A.; Reid, I. AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Brisbane, Australia, 21–25 May 2018; pp. 5882–5889. [CrossRef]
23. Ehsani, K.; Mottaghi, R.; Farhadi, A. SeGAN: Segmenting and Generating the Invisible. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6144–6153. [CrossRef]
24. Follmann, P.; König, R. Oriented Boxes for Accurate Instance Segmentation. *arXiv* **2019**, arXiv:1911.07732.
25. Wei, A.H.; Chen, B.Y. Robotic object recognition and grasping with a natural background. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 42–51. [CrossRef]
26. Zhang, J.; Li, M.; Feng, Y.; Yang, C. Robotic grasp detection based on image processing and random forest. *Multimed. Tools Appl.* **2020**, *79*, 2427–2446. [CrossRef]
27. Huang, Y.J.; Lee, F.F. An automatic machine vision-guided grasping system for Phalaenopsis tissue culture plantlets. *Comput. Electron. Agric.* **2010**, *70*, 42–51. [CrossRef]
28. Yang, L.; Wei, Y.Z.; He, Y.; Sun, W.; Huang, Z.; Huang, H.; Fan, H. iShape: A First Step Towards Irregular Shape Instance Segmentation. *arXiv* **2021**, arXiv:2109.15068.
29. Lin, T.Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014, Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5—12 September 2014*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2015; Volume 8693, pp. 740–755. [CrossRef]
30. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223. [CrossRef]
31. Kirillov, A.; He, K.; Girshick, R.; Rother, C.; Dollár, P. Panoptic Segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9396–9405. [CrossRef]
32. Caldera, S.; Rassau, A.; Chai, D. Review of deep learning methods in robotic grasp detection. *Multimodal Technol. Interact.* **2018**, *2*, 57. [CrossRef]
33. Lenz, I.; Lee, H.; Saxena, A. Deep Learning for Detecting Robotic Grasps. *Int. J. Rob. Res.* **2015**, *34*, 705–724. [CrossRef]
34. Le, Q.V.; Kamm, D.; Kara, A.F.; Ng, A.Y. Learning to grasp objects with multiple contact points. In Proceedings of the IEEE International Conference on Robotics and Automation, Anchorage, Alaska, 3–8 May 2010; pp. 5062–5069. [CrossRef]
35. He, K.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
36. Jiang, Y.; Moseson, S.; Saxena, A. Efficient Grasping from RGBD Images: Learning using a new Rectangle Representation. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3304–3311. [CrossRef]