Image Segmentation Algorithms Composition for Obtaining Accurate Masks of Tomato Leaf Instances

Ivan Zhuravlev[0009−0009−0604−9239] and Andrey Makarenko[0000−0002−3875−7549]

V.A. Trapeznikov Institute of control sciences of russian academy of sciences, Moscow, Russia
{zhursvlevy,avm.science✉}@mail.ru

Abstract. Large agro-industrial complexes are interested in deep automation of the yields control processes to reduce costs caused by errors or a shortage of qualified personnel. Existing approaches solve problems such as yield assessment or plant pathologies detection, but they cannot properly quantify the volume of plant biomass or the diseased area. One of the reasons for this limitation is the poor quality of masks of object instances formed in machine vision systems. This occurs because of Mask R-CNN architecture, which is usually used in the computer vision. In this paper, we propose an algorithms composition for obtaining accurate masks of objects in task of segmentation of tomato leaf instances in images collected in difficult conditions of industrial greenhouses. The use of Mask R-CNN combined with CascadePSP neural network algorithm increased the average IoU by 1.194% compared to "pure" Mask R-CNN on images with complex object-like background.

Keywords: Machine Vision · Video Analytics · Instance Segmentation · Deep Learning · Crop Production.

1 Introduction

Agriculture is an important part of the economy of any country. According to statistics, 55% of the percent of food consumed by humans is vegetables [6]. To provide the population with plant products, industrial greenhouses are being created for growing various types of vegetables and fruits. However, like any living organism, plants need regular care: watering, fertilizers, treatment of diseases. Basically, the analysis of the yield assessment is provided by people manually, and therefore takes a lot of time due to the scale of industrial greenhouses and does not allow timely obtaining complete information about the health of plants, and the identification of plant growth abnormalities and their classification is often carried out by non-specialists, which usually leads to incorrect conclusions about the plant health. Due to these factors, according to the UN [2], about 50% of the crop is lost annually.

Since the process of analyzing the state of crops is monotonous human labor, there is a need to automate this process. Agro-industrial companies are

interested in automatic plant development control systems, as this will help to increase the speed of the crops condition analysis and at the same time reduce the level of errors made in the process of manual analysis. There are various approaches to automating plant health monitoring, in particular, algorithms based on computer vision have been widely used to detect and classify plant pathologies. Intelligent recognition systems are an advantageous solution, since these systems allow visual analysis without the direct involvement of specialists in this process. Thus, machine vision systems not only reduce the time and financial costs of companies, but also reduce the number of errors that are made in the process of manual analysis.

At the moment, computer vision algorithms allow solving problems of detection or classification plant pathologies or calculating yields [12], however, these methods only establish the presence of pathologies, and fruit counting does not provide information about the state of the crops themselves, namely, there is no way to quantify the increase in plant biomass or the relative area of the diseased regions. For visual plant analysis, it is necessary to obtain images of instances of its components. Instance segmentation algorithms, in particular solutions based on Mask R-CNN [9], solve with this task nicely.

However, the instance masks of objects that generate these algorithms have a low resolution. This is critical in tasks such as assessing plant biomass growth and estimating the diseased area, where it is extremely important to obtain accurate information about the shape of objects. This is critical in the agricultural industry. In particular, tomato leaves have a complex shape and the use of Mask R-CNN without additional post-processing introduces a bias in the measurement of the total area of plant leaves or the proportion of diseased regions. In addition to the complex border of the leaf, in real shooting conditions, the leaves are arranged very tightly, the instances overlap each other, and the shooting scene has a uniform color palette, which is why at the moment there are no approaches that solve the problem of assessing plant biomass growth and assessing the affected area.

In this paper, we propose a image segmentation algorithms composition that allows us to obtain accurate instance masks of objects of complex shape, allowing us to solve the problem described above. For this study, images of tomato leaves were collected in difficult conditions of industrial greenhouses. We compared the quality of segmentation with "pure" Mask R - CNN, Mask R - CNN with increased resolution of the mask decoder output and Mask R-CNN with subsequent processing of the model outputs using CascadePSP [7], as well as an algorithm for false positives suppression of the model. Studies have shown that the use of this composition of algorithms gave an increase in the quality of object masks by 1.194% according to the IoU metric averaged over all instances of tomato leaves. Such an improvement characterizes a significant increase in the accuracy of segmentation masks at the border of object instances both in area and shape, which makes it possible to use the proposed method as a basic element in solving problems of assessing the increase in plant biomass and the diseased plant area.

2 Related Work

2.1 Instance Segmentation

At the moment, there are many algorithms of instance segmentation [8]. These ones are divided into two groups: one-stage and two-stage. Among the one-stage methods, the most popular are: YOLACT [5], which generates masks of object instances from a linear combination of mask "prototypes" and is able to work in real-time mode; SOLO [16] use "instance categories" assigning labels to pixels based on the position and size of the object instance; YOLO, which is a SOTA in real-time detection, the latest implementation of which received a branch for instance segmentation. However, one-stage methods are inferior in segmentation quality to two-stage methods.

Among the two-stage ones, one can single out the most popular Mask R-CNN [9], using the so-called region proposal network (RPN), which predicts the objectness score in a specific area of the image which are named region of interest (RoI). The vast majority of modern approaches are based on this architecture. PointRend [10] generates masks of higher resolution than Mask R-CNN (224×224), but requires more data to obtain high-quality masks. Our work does not require real-time processing, so we will give preference to two-stage methods. We choose Mask R-CNN as the basic instance segmentation model.

2.2 High-Resolution Segmentation

Most of the existing high-quality segmentation algorithms were created for the task of semantic segmentation. The main works in this field are [18], [11], [15] . Separately, we highlight the work [7], which offers the CascadePSP algorithm. The authors use a specific approach to model training: the model is trained using an input image and a coarse semantic mask to obtain a mask of increased quality. At the same time, the authors claim that the model does not need additional training and works on any data "out of the box". In our work, we adapt this model to the task of instance segmentation.

2.3 Agricultural Industry

Most of the solutions that exist today are aimed at detecting and classifying plant diseases based on photodata [12]. Thus, in the works [3], [1], the possibility of disease classification based on convolutional networks is investigated. The proposed approaches show high classification accuracy (more than 90% of correct answers), however, these experiments were carried out on datasets containing laboratory data: the leaves are in one copy on a clean background under good lighting. In real shooting conditions, such algorithms degrade greatly due to changing lighting, complexity and diversity of perspectives, the presence of a dense cluster of instances, occlusions. There are also more complex methods for solving the problem of plant disease detection that are tested in real conditions. In the work [14], Faster-RCNN is used as a detector of diseased areas with

subsequent classification of pathologies in bananas. The authors of the article [13] use Mask-RCNN to segment instances of grape fruits to determine the degree of maturity of berries. [4] also uses Mask-RCNN to segment broccoli inflorescence and determine the presence and type of pathology.

These methods make it possible to establish the presence of pathologies or to calculate yields but some of them have been studied to work in favorable simple conditions and are not suitable for use in production, while others, more "advanced" methods based on detection and segmentation algorithms are not able to quantify the proportion of affected plant diseases or biomass growth. The disadvantage of the "advanced" methods is due to the fact that the authors use Mask R-CNN, which generates low-resolution masks (28×28 pixels). This is critical in the task of estimating the growth of biomass or the affected area, since the constituent plants often have a complex shape. In particular, tomato leaves have a complex border relief, so masks of such leaves are of very poor quality. In this paper, we propose a method for obtaining masks of high-quality object instances based on improving the masks generated by Mask R-CNN using CasadePSP and an additional false positive suppression algorithm applied to the problem of segmentation of tomato leaf instances.

3 Dataset

3.1 Agricultural Industry

We have a database of more than 100 thousand images of more than 2,400 high-resolution tomato plants captured in difficult conditions of industrial greenhouses. Data collection took place between October 2020 and January 2022, which corresponds to several cycles of plant development. The process and shooting conditions correspond to the conditions of application of the developed algorithms for wearable cameras, stationary and mobile, installed on a greenhouse robot. Thus, the shooting distance varied from 15 cm to 60 cm. The database contains a description of each image, which includes information about more than 40 pathologies discernible in this image, part of the tomato and the address of the plant. This dataset is designed to build a complex of algorithms for automating the analysis of the state of tomato crops, including for assessing the increase in plant biomass, as well as assessing the area of damage to branches, leaves, petioles, stems and fruits. But, as it was said in the introduction, the main problem of solving such a problem is the very low quality of masks generated by Mask R-CNN-like algorithms. This problem is most pronounced on objects with a very relief contour. Such objects are tomato leaves. In addition to the complex border of the sheet (Fig. 1) the task is also complicated by the uniform green color palette of the entire image (the most striking examples are (a), (b)); an object-like background (b); a large cluster of leaves and branches (d); heavily overlapped or blurred objects (d), different times of day, in in particular, night shooting under artificial lighting (c). This variability and complexity of the data characterizes the actual shooting conditions and distinguishes this dataset from

others used in the works [3], [1]. As a rule, it is not possible to create more favorable angles or conditions for shooting plants, therefore, the assembled dataset most accurately reflects the process of conditions under which the assessment of the state of plants by machine vision methods takes place.

Fig. 1. Data examples.

3.2 Data Markup Methodology

Before the direct construction of the algorithms composition, there was a problem of the data markup methodology. How can the annotator determine which objects are the background of the image, and which are subject to markup. A variant was considered in which all the leaves that are in focus and do not overlap with other objects are marked, but in this case problems may arise in the convergence of segmentation algorithms, since semantically the same object is considered as a target object in one case, and as a background in the other. The second option was to mark up all the non-blurred leaves, but there was a problem in formalizing this process: the degree of blurring varies greatly on the collected data, as a result of which the annotator will not be able to objectively evaluate this parameter. It was decided to mark leaves whose borders do not merge with the background and a person is able to "mentally" mark the boundaries of the leaf. The "labelme" program was used as a tool, which provides the ability to save the markup of objects in the form of coordinates of a polygon bounding a leaf. This program supports the ability to create multiple polygons per instance and combine them into groups, which makes the markup more flexible.

3.3 Data Sampling

A total of 2000 images of leaves were posted. It is worth noting that on different parts of the tomato, the leaves have different shapes, sizes and growth densities. So, in the images where the subject of the shooting were branches, the leaves

are much smaller and more densely arranged than in the images where there was a large leaf in the center of the frame. At the same time, different parts of the plant may have different pathologies. On the branches, the disease "leaf proliferation" is more pronounced, while on the lower level of the plant, large old leaves are susceptible to diseases that manifest themselves in the form of dry leaves. To avoid biases in the data, we made a uniform selection of branches and leaves from several levels of the plant with a volume of 1000 images for manual annotation.

3.4 Semi-Automatic Annotation

After marking up 1000 images, it turned out that on average a person spent about 7 minutes on a photo. In the case of images with a large density of leaves (about 20 items), it could take up to 15-20 minutes for the annotator to process a photo. In this regard, a way to simplify data annotation was proposed. It consisted in pre-learning the composition of algorithms and obtaining rough masks of leaf instances. Preliminary results showed that there is a significant acceleration of marking in semi-automatic mode, especially if there are many small leaves in the image. The quantitative study of the increase in the speed and quality of markup is the subject of the following studies for us, and a more detailed annotation process is described in the section 6. In semi-automatic mode, the remaining part of the images (1000 photos) was marked up, while the sampling was carried out in such a way that more large dry leaves of complex shape were present in the images, since in the first sample, after analyzing the area of the marked objects, a strong imbalance towards small young leaves with a convex smooth border was shown, as indicated by the median area value of 32500 and the value of the 75th quantile (fig. 2).

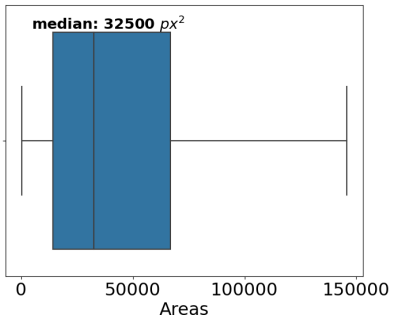

Fig. 2. Distribution of leaves over the bounding box areas.

3.5 Preprocessing

Images are reduced to a single size of 1024×1024 pixels while maintaining the aspect ratio. This image resolution was chosen because we were looking for a trade-off between the computational resources expended and the image detail. To speed up the the mini-batch loading, polygonal annotation is converted into binary masks. In addition to masks, the coordinates of the bounding boxes are also calculated. At the time of the mini-batch formation, the pixels of the image are normalized to the segment [0; 1]. Additionally, standard augmentation is applied in the form of random rotations and mirror reflections to ensure the invariance of the algorithm to the rotation of objects.

4 Method

As mentioned earlier, Mask R-CNN produces low-resolution masks (28×28 or 14×14 pixels). With an increase in resolution, the quality of masks does not lead to a significant improvement, while the time for calculating the outputs of the mask decoder increases. In our work, we also conducted experiments with increasing the size of feature maps at the output of the mask decoder and did not get a significant improvement in quality. In some tasks, high-quality masks are not required, but in the case of segmentation of objects of complex shape, where the area of the object plays a key role in the subsequent stages of solving the problem, this is a serious limitation of the algorithm. In the problem of tomato leafs segmentation under consideration, with subsequent assessment of biomass growth and diseased area, much higher accuracy is required than that which can be obtained using a "pure" Mask R-CNN.

However, we note the advantage of this architecture. Mask R-CNN is an extension of Faster-CNN, a two-stage detector that uses the region proposal network (RPN) to generate regions of interest, which provides high quality object detection in the image. One-stage methods do not use RPN, so they lose in quality. In addition, Mask R-CNN has a "modular" architecture and can easily adapt to a specific task. Therefore, the choice of most researchers falls on this neural network. The listed properties inspired us to keep this architecture in use, but to introduce additional post-processing.

4.1 Mask R-CNN Settings

The convolutional network ResNet-50 in combination with the feature pyramid network (FPN) was chosen as the backbone. It is known that the deep layers of the convolutional neural network (CNN) contain semantic features, but lose information about the details of the object, for example, about the boundaries. In contrast, the upper layers of CNN contain more general features, such as lines or angles. FPN allows you to take into account both the detailed characteristics of the object, as well as semantic, which is important when segmenting instances of objects. Anchors were generated at the following scales: $[64, 128, 256, 512, 1024]$

with aspect ratio $[1:2, 1:1, 2:1]$. Such variability of scales is necessary due to the large dispersion of leaf areas. The configuration of the heads of the regions of interest was chosen the same as described in the original article, except for the number of classes, since there is only one class in our problem. To experiment with increasing the output resolution of the mask decoder, we used 4 deconvolution layers [17] with ReLU activations. The use of more layers did not lead to quality increase. The predicted masks of objects are interpolated to the size of the original image 1024×1024 pixels.

4.2 Masks Refinement

Our goal is to perform a transformation that would receive coarse masks of the object in the image as input and give out an improved quality mask at the output. CascadePSP [7] implements such a transformation. Initially, this algorithm was developed to improve masks in the semantic segmentation problem for subtask named scene parsing. This means that the single mask contains all objects that belong to the same class, both during model training and inference.

We assumed that this model can be adapted to the instance segmentation problem. We trained the model on semantic masks of objects, but during the inference, we submitted masks of instances to the input. Experiments have shown that this approach is working and object masks are significantly improved, in particular, on large objects of complex shape. The network architecture and loss functions are taken the same as in the original article. During the output of the model, two steps are performed to improve the mask: global and local. The first step takes the mask and the image as an input in its entirety and produces an initial improvement of the mask. During the second step, the image is fed in parts: regions of 224×224 pixels are cut out on the image and the mask obtained in the first step, then the resulting improved masks are concatenated into the final version, which is the output of the model. Note that the authors of the article claimed that this algorithm does not need to be retrained on new data, but early experiments demonstrated degradation of the quality of masks on our specific task. More specifically, the improved mask captured the petiole, which is unacceptable.

4.3 Suppression of False Positive Hypotheses

During the testing of the algorithm composition, a visual analysis was made and a problem was found related to the filtering of hypotheses (network outputs) by the non maximum suppression (NMS) algorithm. Tomato leaves have a complex shape, part of the leaf is similar to the whole leaf, twisting forms fictitious boundaries, as a result of which the segmentation algorithm recognizes part of the leaf as a separate instance (false positive hypothesis). An original algorithm "false positive suppression" (FPS) is proposed.

Let $S_i = (B_i, M_i), i = \overline{1, N}$, where N is the number of hypotheses, B_i, M_i is the bounding box and the instance mask. Let a hypothesis with a larger and smaller area of bounding box be found among a pair of hypotheses S_i and S_j,

then we denote the corresponding bounding box and masks as $B_{\min}$, $B_{\max}$, $M_{\min}$, $M_{\max}$. Checking whether the smaller hypothesis is a false positive is performed as follows. The metric $IoU(S_1, S_2)iscalculated$:

$$IoU(S_1, S_2) = IoU(M_{\max}^{crop}, M_{\min}), \tag{1}$$

where $M_{\max}^{crop}$ is calculated as follows:

$$M_{\max}^{crop} = crop\_mask(M_{\max}, B_{\min}). \tag{2}$$

If the metric (1) exceeds the specified threshold, then the smaller hypothesis is considered a false positive and is removed from the list of hypotheses. In our problem, we took a threshold of 0.5. The function (2) cuts out on the mask of the larger hypothesis the area occupied by the bounding box of the smaller hypothesis.

4.4 Algorithms Composition Inference

Schematically, the sequence of obtaining accurate masks is shown in Fig. 3. The part that is used to obtain the final results in plant condition monitoring tasks is highlighted in blue, and the proposed modification for solving the problem of segmentation of objects of complex shape is highlighted in red. An image is fed to the Mask R-CNN input, the output of the model contains an bounding boxes, classes of object instances and instance masks filtered by a given threshold of objectness score in the region of interest and the NMS algorithm. The resulting masks, together with the original image, are alternately passed through CascadePSP, then hypotheses with improved masks are processed by the FPS algorithm and at the output we get a set of bounding boxes, classes, as well as masks of high-resolution instances.

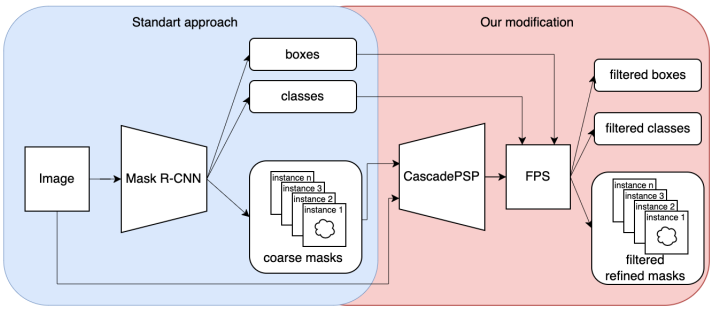

Fig. 3. A scheme for obtaining improved masks of object instances.

4.5 Training Parameters

Mask R-CNN. Due to the small size of the training sample, the R-CNN pre-trained on MS COCO Mask was used. The weights on the first two residual resnet-50 blocks of Mask R-CNN were frozen, since these layers contain the most common features. SGD with an initial learning rate of $5 \cdot 10^{-4}$ with a 10-fold decrease in speed after the 70th epoch was chosen as the optimization method. The total number of epochs is 100. The warmup technique was also used in the first epoch of training. A linear increase in speed was performed over two hundred iterations to the initial learning rate. L_2 regularization with a coefficient of $5 \cdot 10^{-4}$ was also applied. The size of the mini-batch is 8 examples. The volume of the training and validation samples was 80% and 20%, respectively.

CascadePSP. The BIG dataset [7] weights pre-trained on the were used. The configuration of the optimizer and the learning rates are the same as described in the original article. The size of the mini-batch is 8 examples. Before submitting images and semantic coarse masks to the input, a random cutout was made on the image and masks with a size of 224×224 pixels.

5 Experiments

The training results are presented in the table 1. Precision was used as a metric to estimate the proportion of false positives, recall to estimate the proportion of false omissions, and F1-measure, since it is important for us that Mask R-CNN detects all instances of tomato leaves. 4 models were compared: basic Mask R-CNN and one with increased mask resolution up to 112×112; Mask R-CNN + CascadePSP and with a FPS filter. To evaluate the quality of masks, the IoU metric was used, averaged over all detected leaf instances. The results show an improvement in the average IoU by 1.194% with the best confidence threshold of the network $t_{objectness} = 0.8$ for the latest model in the table compared to the basic one.

Table 1. Models comparison

Model	**precision**	**recall**	**F1**	**average IoU**	**ms/image**
Mask R-CNN	0.8656	0.8349	0.8500	0.8896	**202**
Mask R-CNN 112×112	0.8541	0.8563	0.8552	0.8912	211
Mask R-CNN + CacadePSP	0.8675	0.8367	0.8518	**0.9018**	5540
Mask R-CNN + CascadePSP + FPS	**0.8737**	0.8342	0.8535	**0.9016**	5663

Training and inference of both Mask R-CNN and cascadePSP is carried out by parallelizing calculations on a single Nvidia Titan RTX 24GB GPU using NVIDIA CUDA. The inference time of the model is greatly increased when Mask R-CNN followed by cascadePSP is used (the last column of the table 1). This is due to the repeated passing of a single instance mask through CascadePSP, while the output of each iteration of the improvement produces caching, which affects

the amount of video memory used ≈3GB per instance. At the moment, Cascade PSP does not support mini-batch calculations during model inference, however, parallelization of the algorithm by object instances or video streams is possible, which will reduce the total inference time. In addition, a real-time algorithm is not required for this range of tasks, so when developing an approach, we give preference primarily to the quality of the model. Our experiments demonstrate that the task is solvable, so optimizing the algorithm in terms of inference time and memory usage will be the next step.

In order to verify the results obtained, a Student's t-test was conducted on a sample of the average IoU calculated from the outputs of the basic model and improved one. The null hypothesis was that there are no differences in the average values of IoU, the alternative is that there are. The significance level was chosen 10^{-4}. The table 2 shows the p-value obtained during the comparison of models. At the selected significance level, simply adding mask decoder layers does not give statistically significant results (first line), however, an improvement in the model is observed when comparing Mask R-CNN in conjunction with CascadePSP.

Table 2. t-test for average IoU

Base model	**Improved model**	**p-value**
Mask R-CNN	Mask R-CNN 112×112	0.543
Mask R-CNN	Mask R-CNN + CascadePSP	10^{-6}

A visual comparative analysis is shown in Fig. 4. Columns (a), (b), (c), (d) – the original image, markup, Mask R-CNN, outputs of the algorithms composition respectively. The top two rows demonstrate how the quality of the mask changes in large leaves with a complex border, almost pixel-by-pixel accuracy is observed. Note that the difference in the quality of small masks is poorly traced (third row), which is logical, since small leaves have a convex shape and a fairly simple border, which potentially makes it possible to use Mask R-CNN for objects of this size without additional processing using CascadePSP, however, the research data relate to the optimization of the model in this work are not carried out. In the last row, you can see how the FPS algorithm works (columns (c) and (d)). Part of the leaf was recognized as a separate instance and was filtered. There are also some inaccuracies of the mask in the last row, which are associated with a difference in the brightness of the areas of the sheet. With the help of additional augmentation by varying brightness, as well as the expansion of the dataset, such side effects will disappear.

The question may arise whether the improvement of masks is significant from the point of view of application to the task of determining the increase in plant biomass and assessing the diseased area. The answer is yes, because often dangerous pathologies, such as cancer in the early stages, manifest themselves at the borders of the leaf. These areas are cut off during rough segmentation, so there is a risk of detecting pathologies only at the moment when the area of

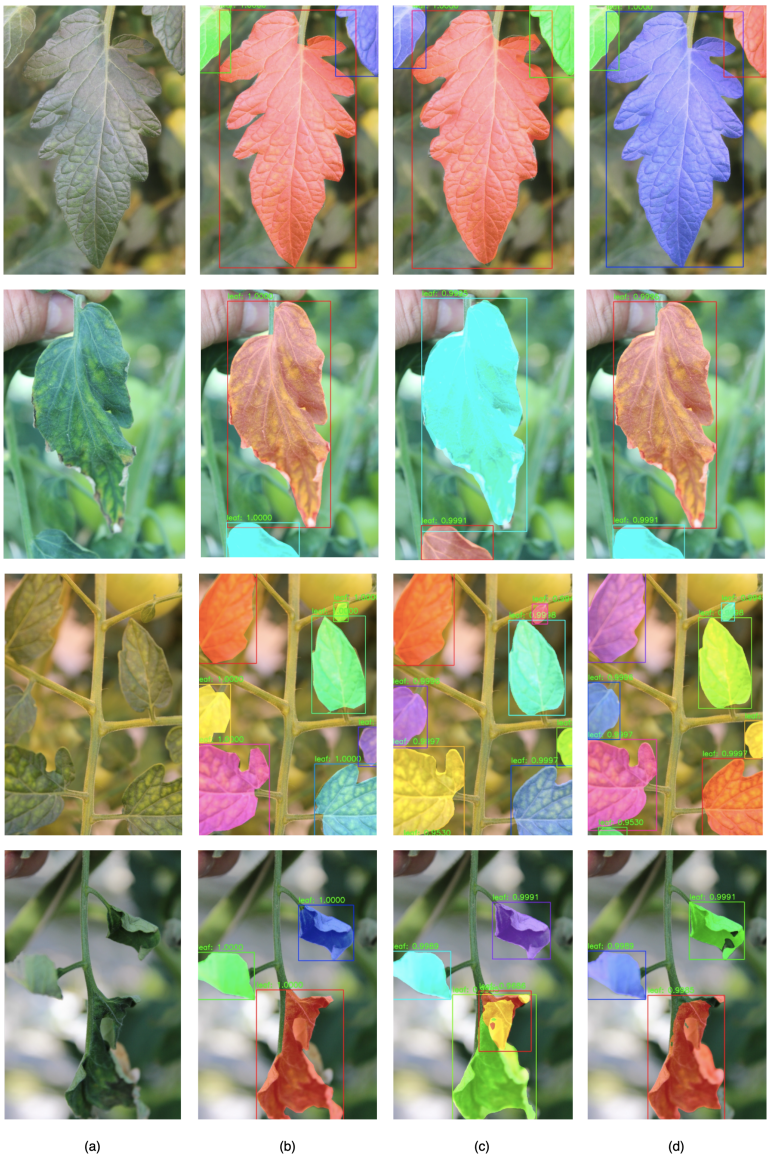

Fig. 4. Model inference examples. (a) – image, (b) – markup, (c) – Mask R-CNN, (d) – Mask R-CNN + CascadePSP + FPS.

lesions acquires a large scale. This problem is most clearly demonstrated in the second row in Fig. 4. "Clean" Mask R-CNN captures the border along with the background, or, conversely, does not cover the affected area. The algorithms composition segments with high accuracy.

It should be noticed that at first glance there is a problem: for both tasks it is necessary to know the distances to the objects in order to calculate their actual dimensions, but the distance to the shooting plane is unknown to us with the required accuracy. In many cases, it is not the actual volume that is important, but its increase relative to the previous biomass assessment session. It is the dynamics of growth that often turns out to be a key factor in assessing the intensity of plant development. At the same time, if the shooting is carried out by a mobile robot then each plant is evaluated (i.e. a digital 4D map of the greenhouse is being built) with yield forecasts already being made based on this data. If the shooting is carried out "by hand, by an agronomist" or by means of stationary cameras, then mathematical statistical criteria are applied: for the observed plants, an average estimate and confidence intervals for the greenhouse block as a whole are output.

Thus, we come to the conclusion that this algorithm can be used as a basic element of systems for assessing yield and diagnosing plant pathologies. Our algorithm has been investigated in relation to tomatoes but can be transferred to other plants. The use of the improving module makes it possible to estimate the area and shape of plant parts with high accuracy.

6 Semi-Automatic Data Markup

Since the deadlines for solving the problem were short and the financial possibilities were small, we needed optimization in data markup. 7 minutes to mark up one image is a lot. The monotony of the process of marking complex objects for a long time, as shown by the visual analysis of the first 1000 marked photos, led to errors in the markup due to the human factor (fatigue or inattention of the marketer). After 1000 images were marked up completely manually, it was proposed to train the composition of algorithms on the available data. Our assumptions were that even poorly trained algorithms would simplify the markup process. The resulting rough masks at the Cascade PSP output can be converted to a data markup format and corrections can be made manually. Intuition suggests that it is easier to correct the inaccuracies of masks, remove noise or add missing instances than to make markup "from scratch".

Fig. 5 shows the process of obtaining preliminary data markup. A selection of data is made from the database. Images are passed through Mask R-CNN and CascadePSP. The resulting masks of object instances must be translated into markup format. It implies storing masks in the form of coordinates of polygon faces. If we use the transformation of the mask into an array of coordinates directly, we will get a huge number of points, which on the contrary will complicate the process of additional markup. Therefore, it is necessary to reduce the number of coordinates in such a way that the markup is comfortable, but at the same time not to lose the quality of the preliminary markup.

We analyzed the distribution of the number of polygon points set during manual marking and the area of instances (Fig. 6). Our assumption was that the best number of points should correspond to the average number of points

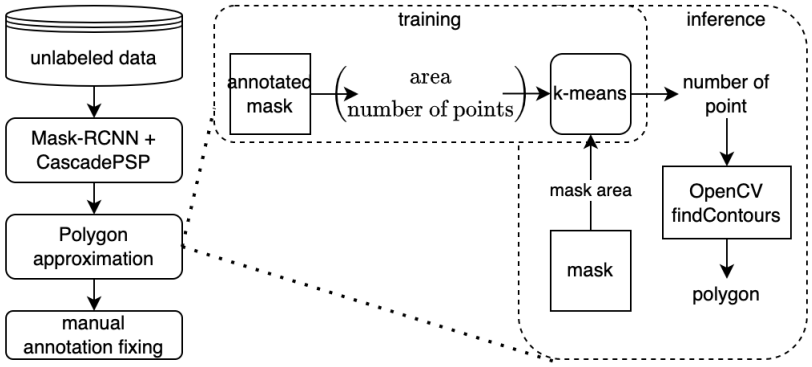

Fig. 5. Scheme for obtaining preliminary data markup.

that falls on an instance of a particular area. We used the k-means method to cluster instances by the number of coordinates and areas.

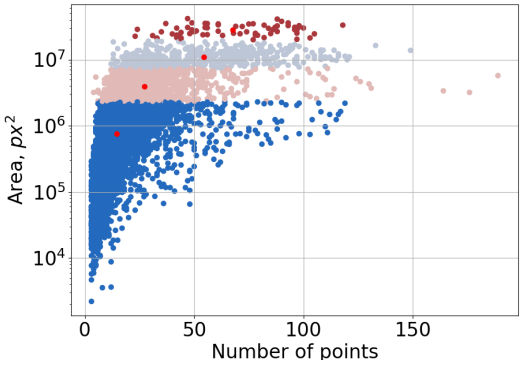

Fig. 6. Joint distribution of the instance mask areas and the number of polygon points during marking.

Thus, the process of approximation of polygons do in the following way. The area of the instance mask is calculated, the nearest cluster center of the distribution of areas and the number of points is located, and the average number of points for this cluster is matched to this instance mask. Then, by means of OpenCV, the mask is converted into a polygon. We chose to split the distribution into 4 clusters, and also tripled the number of points when converting the mask into a polygon, since already at this stage the outputs of CascadePSP turned out to be quite accurate. Examples of the obtained preliminary markup are shown in Fig. 7.

Fig. 7. Examples of the received preliminary markup.

To quantify the effectiveness of markup by the time of its execution, it is necessary to conduct testing with measuring the time spent by the markup without using this approach and using one. These studies will be carried out in subsequent works.

7 Conclusion

In this paper, a segmentation algorithms composition was proposed to obtain accurate instance masks of complex objects in relation to the problem of segmentation of tomato leaves. The main feature of the method was the use of the CascadePSP neural network algorithm as a postprocessing of the Mask R-CNN segmenter output. A statistically significant increase in the quality of masks of tomato leaf instances by 1.194% was obtained, while the algorithm achieved such an increase in quality on a small dataset with a volume of 2000 images. In addition, an algorithm for suppressing false positive hypotheses was proposed to filter Mask R-CNN outputs for objects of complex shape. An algorithm for obtaining preliminary data markup was also proposed and tested using the composition of mask improvement algorithms proposed in this paper, trained on a small dataset.

The results of this work demonstrated that the solution of the problem of assessing the increase in biomass and the area affected by pathologies is possible and this approach can be used as a basic part of the plant development monitoring system. Our algorithm can be adapted to different types of plants. The use of the mask refinement module makes it possible to estimate the area and shape of plant parts with high accuracy, and the shooting conditions in which the data were collected make it possible to apply this method using both wearable cameras and stationary ones. In addition, mobile cameras installed on the greenhouse robot are also supported. In the future, it is planned to optimize the output time of the model, the efficiency and speed of marking, as well as the introduction the algorithms composition into the systems for monitoring the condition of crops of agro-industrial companies.

The authors thank the anonymous referees for their helpful comments.

References

1. and, S.J.R.: Plant disease detection using transfer learning in precision agriculture. AMBIENT SCIENCE **9**(3), 34–39 (nov 2022). https://doi.org/10.21276/ambi.2022.09.3.ta02
2. Arunnehru, J., Vidhyasagar, B.S., Basha, H.A.: Plant leaf diseases recognition using convolutional neural network and transfer learning. In: Lecture Notes in Electrical Engineering, pp. 221–229. Springer Singapore (2020). https://doi.org/10.1007/978-981-15-2612-1_21
3. Arunnehru, J., Vidhyasagar, B., Anwar Basha, H.: Plant leaf diseases recognition using convolutional neural network and transfer learning. In: International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2019. pp. 221–229. Springer (2020). https://doi.org/10.1007/978-981-15-2612-1_21
4. Blok, P.M., Kootstra, G., Elghor, H.E., Diallo, B., van Evert, F.K., van Henten, E.J.: Active learning with MaskAL reduces annotation effort for training mask r-CNN on a broccoli dataset with visually similar classes. Computers and Electronics in Agriculture **197**, 106917 (jun 2022). https://doi.org/10.1016/j.compag.2022.106917
5. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: Real-time instance segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (oct 2019). https://doi.org/10.1109/iccv.2019.00925
6. Cassidy, E.S., West, P.C., Gerber, J.S., Foley, J.A.: Redefining agricultural yields: from tonnes to people nourished per hectare. Environmental Research Letters **8**(3), 034015 (aug 2013). https://doi.org/10.1088/1748-9326/8/3/034015
7. Cheng, H.K., Chung, J., Tai, Y.W., Tang, C.K.: CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020). https://doi.org/10.1109/cvpr42600.2020.00891
8. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. International Journal of Multimedia Information Retrieval **9**(3), 171–189 (jul 2020). https://doi.org/10.1007/s13735-020-00195-x
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR **abs/1703.06870** (2017), http://arxiv.org/abs/1703.06870
10. Kirillov, A., Wu, Y., He, K., Girshick, R.: PointRend: Image segmentation as rendering. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2020). https://doi.org/10.1109/cvpr42600.2020.00982
11. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jul 2017). https://doi.org/10.1109/cvpr.2017.549
12. Rakhmatulin, I., Kamilaris, A., Andreasen, C.: Deep neural networks to detect weeds from crops in agricultural environments in real-time: A review. Remote Sensing **13**(21), 4486 (nov 2021). https://doi.org/10.3390/rs13214486
13. Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S.: Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Computers and Electronics in Agriculture **170**, 105247 (mar 2020). https://doi.org/10.1016/j.compag.2020.105247

14. Selvaraj, M.G., Vergara, A., Ruiz, H., Safari, N., Elayabalan, S., Ocimati, W., Blomme, G.: AI-powered banana diseases and pest detection. Plant Methods **15**(1) (aug 2019). https://doi.org/10.1186/s13007-019-0475-z
15. Shen, T., Zhang, Y., Qi, L., Kuen, J., Xie, X., Wu, J., Lin, Z., Jia, J.: High quality segmentation for ultra high-resolution images. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2022). https://doi.org/10.1109/cvpr52688.2022.00137
16. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: SOLO: Segmenting objects by locations. In: Computer Vision – ECCV 2020, pp. 649–665. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-58523-5_38
17. Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R.: Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE (jun 2010). https://doi.org/10.1109/cvpr.2010.5539957
18. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. In: Computer Vision – ECCV 2018, pp. 418–434. Springer International Publishing (2018). https://doi.org/10.1007/978-3-030-01219-9_25