RS-Net: Robust Segmentation of Green Overlapped Apples

Weikuan Jia[1, 2], Zhonghua Zhang[1], Wenjiang Shao[1], Ze Ji[3], Sujuan Hou[1, 4]

Abstract:
Fruit detection and segmentation will be essential for future agronomic management, with applications in yield estimation, growth monitoring, intelligent picking, disease detection and etc. In order to more accurately and efficiently realize the recognition and segmentation of apples in natural orchards, a robust segmentation net (RS-Net) framework specially developed for fruit production is proposed. This model was improved for the more challenging problem which segments the overlapped apples from the monochromatic background regardless of various corruptions. The method extends Mask R-CNN by embedding an attention mechanism for focusing more on the informative pixels but also suppressing the noise caused by adverse factors (occlusions, overlaps, etc.), which could be more suitable and robust for operating in complex natural environment. Specifically, the Gaussian non-local attention mechanism is transplanted into Mask R-CNN for refining the semantic features generated continuously by Residual Network (ResNet) and Feature Pyramid Network (FPN), then the model forward processing based on the balanced feature levels and finally segments the regions where the apples are located. Experimental results verify the hypothesis of current work and show that the proposed method outperforms other start-of-the-art detection and segmentation models, the AP box and AP mask metric values have reached 85.6% and 86.2% in a reasonable run time, respectively, which can meet the precision and robustness of vision system in agronomic management.

Key Words:
Robust Segmentation, Overlapped Apples, Mask R-CNN, RS-Net

✉
Weikuan Jia
jwk_1982@163.com
Sujuan Hou
hsj1985@126.com

Abbreviations

AP	Average precision %
AR	Average recall %
BFP	Balanced feature pyramid
CHT	Circular hough transform
CNN	Convolutional neural networks
FCN	Fully convolution network
FN	False negative
FPN	Feature pyramid network
IoU	Intersection of union
MLP	Multiscale multilayered perceptron
NMS	Non-maximum suppression
R-CNN	Region-based convolutional network
ResNet	Residual network
RoI	Region of interests
RPN	Region proposal network
RS-Net	Robust segmentation net
TP	True positive
WS	Watershed segmentation

Introduction

To supply the nutrition and health needs of the growing population around the world, a major challenge in agricultural communities is to find innovative ways to increase the production of fruits and vegetables (Siegel et al., 2014), especially in the context of rising farming costs and the shortage of skilled labor. Efficient and sustainable agronomic management is one of the effective ways to alleviate this situation, which is required to reduce economic and environmental costs while increasing orchard productivity. Recently, advances in technologies such as robotics and computers provide farmers with means to increase agricultural production in an efficient and sustainable way (Underwood et al., 2016). In addition, these new technologies has been widely applied in the optimization of processes in agronomic management such as irrigation, fertilization, pruning, thinning and deinsectization (Auat Cheein and Carelli, 2013; Bargoti and Underwood, 2017b), through the detection and quantification of fruit distribution in canopy, farmers can obtain valuable

information and provide reference for optimizing these processes, which will significantly facilitate the spatial and temporal management of agricultural production.

Among the many links to realize efficient and sustainable agronomic management, vision system as the most fundamental yet important section, used to parse the specified targets from the complex and diverse scenes, has been widely used in many practical applications, such as crop yield estimation (Koirala el al., 2019a), growth monitoring (Fu et al., 2020b), intelligent picking (Bac et al., 2015), disease detection (Zhang et al., 2019), and so on. Design of vision system with the goal of rapid positioning and accurate segmentation will significantly affect the real-time and reliability of these intelligent agriculture applications. However, there are many different types of interference under natural conditions, such as various scales, occlusions, overlaps, illuminations, etc., especially in the monochromatic background, which are all unfavorable to visual system and need to be taken these factors into consideration. Therefore, how to enhance the discriminative ability of vision system regardless of the above interference is crucial and necessary. In this paper, a robust segmentation net framework is specifically designed to segment the overlapped apples from the monochromatic background, which will be more challenging than previous works (Zhang et al., 2016; Jia et al., 2020b).

In recent years, many researchers have been attracted and proposed different methods for the improvement and robustness of detection model in complex orchard scenes. Some methods for detecting used colorspace transformations where the objects of interest stand out, or extraction of features such as shape and texture (Kapach et al., 2012; Liu et al., 2016b; Rong et al., 2012; Jia et al., 2015; Gongal et al., 2015). In most of these solutions based on hand-crafted features, the discriminative information depends partly on developers, not entirely on algorithms themselves, which may not enough to deal with the level of variability and complexity which commonly appear in natural orchards. In addition, some scholars proposed computer vision solutions based on deep learning network architecture (Chen et al., 2017; Jia et al., 2020a; Fu et al., 2020a; Li et al., 2021; Vasconez et al., 2020). Although these methods can deeply mine the characteristics of targets by themselves, some inconspicuous targets are easily disturbed by dominated salient objects and cause wrong judgments. This situation is sharper when recognizing overlapped fruits at monochromatic background, which still cannot meet the needs of real-world application.

Through the analysis of the above problems, the objective of this study linking image processing with agronomic management was to develop a model architecture with strong robustness to segment apples regardless of interference caused by sensors and natural orchard elements. The whole network framework can be divided into three parts: (1) Feature Acquisition, (2) RoIs (Region of Interests) Generation and (3) Results Prediction. Firstly, the pipeline of ‘Feature Acquisition’ also consists of three steps: extraction, fusion and refining, which are respectively performed by residual network (ResNet) (Targ et al., 2016), feature pyramid network (FPN) (Kim et al., 2018) and balanced feature pyramid (BFP) (Pang et al., 2019). The features of each image were extracted by ResNet and fused by FPN successively, which can make different scales caused by diverse factors (occlusion, camera distance and angle, etc.) be all well perceived. Sequentially, BFP strengthen the features from FPN with the embedding of Gaussian non-local attention mechanism, which can retain more semantic information of inconspicuous object by selectively integrating the similar features rather than simple contextual embedding. Then, at the stage of ‘RoIs Generation’, the region proposal network (RPN) (Ren et al., 2017) takes the features refined by BFP as input and outputs a set of rectangular object proposals on original images, each with a score that belongs to

the foreground. Afterwards, the RoI Align layer convert the features inside any valid region of interest into a small feature map with a fixed spatial extent of $H \times W$, where H and W represent the height and width of RoIs respectively. The RoIs with same size are transported into three branches of 'Results Prediction' for class probability, bounding box (bbox) regression and mask generation respectively. Finally, based on the results generated by three branches and combine with them, the model will get the final segmentation results.

It should be noted that the method is more effective and flexible than previous methods which also based on network architecture when dealing with complex and diverse scenes. Specifically, some fruits are inconspicuous or incomplete due to lighting and occlusion. If simple contextual embedding is explored, the semantic information from dominated salient object (e.g. leaves, branches) would harm those inconspicuous objects labeling near the edge. By contrast, the embedding Gaussian non-local attention module selectively aggregates the similar features of inconspicuous objects to highlight their feature representations and avoid the influence of salient objects. In addition, through the way that explicitly take spatial relationships into account, so that images understanding for segmentation could benefit from the whole building long-range dependency. Compared with previous published work, the current work contributes to the development of a solution for vision system in agronomic management by examining the hypothesis that Gaussian non-local attention mechanism can be easily embedded into deep learning based vision model and effectively improve the accuracy and robustness of fruit detection by aggregating the similar features of inconspicuous objects through the image. In general, this study offers at least the following contribution as:

(I) Gaussian non-local attention mechanism is embedded to focus on the informative pixels but also suppress the noise.

(II) The proposed methods outperform the start-of-the-art models in terms of both accuracy and robustness, which could be more suitable for detecting fruits in complex scenes.

(III) Provide valuable reference for practical application of other fruit detection and segmentation methods.

The rest of this paper is organized as follows: Firstly, section 2 briefly outlines the breakthrough of related works and unresolved issues. Next, section 3 introduces image acquisition and related dataset processing and annotations. The detailed improvement of model architecture and whole network's pipeline will be illustrated in section 4. In section 5, the experiment is shown to validate that the method outperforms others from different perspectives including precision, recall and robustness. Finally, section 6 summarizes the characteristics of the proposed method and elaborates the other unsolved problems in this field, which will be the future research directions.

Related work

Design of vision system with the goal of rapid positioning and accurate segmentation is a very challenging task. This is due to various complicated and changeable situations in natural orchards. For example, occlusions and overlaps will lead to incomplete shape features, angle and intensity of illuminations will lead to the indistinct texture features, etc. In the domain of agriculture, earlier work about this used "classical" machine vision techniques, involving detection, classification and segmentation tasks based on hand-crafted features. For example, Ji used SVM classifier to classify and recognize apples, and the recognition rate of bagged apples reached 89%, however, it took

352ms to recognize an image, the recognition efficiency was not enough high and could not meet the real-time requirements (Ji et al., 2012). Tian proposed an optimized graph-based recognition algorithm by utilizing depth images and paired RGB images without extra manual labeling, which achieves both accuracy and speed improvement, but there were obvious defects in the segmentation of overlapping and clustered apples (Tian et al., 2019a). Liu proposed a recognition method for bagged apples based on block classification, watershed algorithm was employed to segment original images into irregular block, and then SVM divided these blocks into fruit blocks and non-fruit blocks, which can restrain the interference of light efficiently (Liu et al., 2018). Rakun used to recognize apples by combining texture and color features, but the bags and drops on the apples would weaken or even change the features and make it difficult to recognize them (Rakun et al., 2011). Bargoti and Underwood proposed a pipeline for mango and apple detection and counting. They used a general-purpose image segmentation approach with two feature learning algorithms—convolutional neural networks (CNN) and multiscale multilayered perceptrons (MLP). Their approaches were designed to include contextual information about how the image data were captured. Circular Hough transform (CHT) and watershed segmentation (WS) algorithms were used to detect and count individual fruits from the pixel-wise fruit segmentation (Bargoti and Underwood, 2017a; Bargoti and Underwood, 2017b). Linker proposed a yield prediction model specifically for night-time apple images. In addition, the classifier trained with images from one dataset was successfully applied to the second dataset, and the same prediction effect as the previous work was achieved (Linker, 2018). Hung demonstrated a generalised multi-scale feature learning approach to multi-class segmentation of tree crops. The segmentation results were applied to the problem of fruit counting and compared against manual counting, which shows a squared correlation coefficient of $R^2 = 0.81$ between the two (Hung et al., 2015). Similarly, there are other methods to realize fruit detection by combining color, texture, shape and other features (Aggelopoulou et al., 2011; Kurtulmus et al., 2011; Wang et al., 2013), these 'classical' machine vision based methods rely heavily on hand-crafted features to refine discriminative information, so that they could not yet comprehensively consider more aspects into account, which will be eliminated in complex real-world environment.

With the gradual maturity of deep learning, it has become prevalent to migrate this novel revolution to various professions for better results. This also has stimulated the development of vision system in precision agricultural field. More recent works draw support from deep learning due to its various flavors and strong adaptability. Gené-Mola used RGB-D cameras to collect geometrical information with color data and adapted Faster R-CNN model for use with five channels input images including color (RGB), depth (D) and range-corrected intensity signal (S). Results show an improvement of 4.46% in F1-score when adding depth and range-corrected intensity channels, which can be concluded that the RGB-D sensors give valuable information for fruit detection (Gené-Mola et al., 2019). Li optimized U-Net with gated and atrous convolution to make the model more suitable for small apple segmentation in monochromatic background, and the recognition time is 0.39 seconds. However, the optimized U-Net still belongs to semantic segmentation model, which can only achieve pixel-level classification instead of instance-level. This results in overlapping and clustered fruits being divided into an area, which is not suitable for fruit counting and picking (Li et al., 2021). Koirala compared six existing deep learning architectures for the task of mango detection, and developed a new architecture named MangoYOLO based on features of YOLOv3 and YOLOv2(tiny) on the design criteria of accuracy

and speed. The MangoYOLO achieves a F1 score of 0.968 with a detection speed of 8 ms, which realizes a good trade-off between speed and accuracy (Koirala et al., 2019b). Tian employed the improved YOLOv3 architecture to detect apples during different growth stages. Images of young apples, expanding apples and ripe apples were initially collected and subsequently augmented. These augmented images were sent into DenseNet for feature extraction.(Tian et al., 2019b). Sa adapted an object detector by using a Faster R-CNN through transfer learning with images obtained from two modalities—color (RGB) and Near-Infrared (NIR). However, this model used vgg16 as the backbone of the whole model pipeline to extract features, the large capacity (more than 500 megabytes) of the model may make it difficult to deploy the model to mobile agriculture devices (Sa el al., 2016). Rahnemoonfar and Sheppard trained a CNN using features on multiple scales based on an Inception-ResNet architecture. The model was trained with synthetic images and tested on real images of tomato plants, reaching 91% of accuracy. However, such network was tested using 128 × 128 pixel images, which may not take into account important features from the fruit due to the low resolution (Rahnemoonfar and Sheppard, 2017). Vasconez tested the effect of two most common CNN detectors (Faster-RCNN with Inception and SSD with MobileNet) in fruits detection, and compared the results of the two models on three fruits datasets (avocado, apple and lemon). Extensive experiments provide insightful analysis of the usability of such technique in fruit counting tasks in groves, which can lead to further improve the decision making process in agricultural practices (Vasconez et al., 2020). Though improved network enhanced the feature propagation and reusability, research data shows that the algorithm would still easily affected by occluded objects.

Although many researches in this field have made breaks through various ways, the above methods could only realize one of the functions between detection and segmentation, Mask R-CNN (He et al., 2019; Wei et al., 2018) provides a framework for prediction of both the bounding box and pixel mask for each object with only adding a mask branch on Faster R-CNN (Ren et al., 2017), which can efficiently eliminate interference caused by overlaps and occlusions. Considerable amount of researches about Mask R-CNN based methods are under study and gain some progress. Jia improved the backbone of Mask R-CNN by combining ResNet with DenseNet, which greatly reduce input parameters and efficiently strengthen feature extraction (Jia et al. 2020b). Yu applied Mask R-CNN for detecting strawberries and get ideal effect in terms of both robustness and universality, particularly for inconspicuous fruits (Yu et al., 2019). Otherwise, many works about attention mechanism (Fu et al., 2019; Wang et al., 2018; Chen et al., 2016) also make great breakthroughs, which can provide references for detecting overlapped apple from monochromatic background. Inspired by these two innovations and combined with the goal of segmenting occluded apples in monochromatic background, the current work proposed an improved model based on Mask R-CNN by embedding the Gaussian non-local attention mechanism for better focusing on the informative pixels but also suppressing the noise.

Dateset Generation

Images acquisition

The images were acquired at the Longwangshan apple production base in Fushan District, Yantai City, Shandong Province (agricultural information technology experimental base of Shandong Normal University) with 6000 × 4000 pixel resolution. Generally, the performance of models based on deep learning relies heavily on datasets, in order to satisfy the diversity of

overlapped fruit recognition and minimize variability in lighting conditions due to direct sunlight or cloud cover, images was taken at multiple directions and multiple time interval (morning, noon and evening). Totally, 268 apple images with different illumination and different amounts were collected. Otherwise, considering that the detection of overlapped apple in the monochromatic background is taken as the research object of this paper, these collected images contains a large proportion of apples occluded by leaves and branches or overlapped by another fruits. Specifically illustrated in the first two rows of Figure 1.

Images annotation and dataset production

Although the aforementioned factors have benn taken into account when collecting photos, some literatures (Dodge et al., 2016; Michaelis et al., 2019) have proved that most standard detection and segmentation models encounter a serious detection loss when images getting corrupted(down to 30-60% of the original), which are inevitable caused by sensor degradation or poor weather and extremely unfavorable for real-world applications of agronomic management. For example, the state-of-the-art segmentation algorithm such as Mask R-CNN failed to segment partial apples when the fog gets thicker(as shown in the third row of Figure 1), even though the apples are still clearly visible to human eyes, which means that the vision system will be a bad alternative of manual labor if the robustness of models cannot get improved. Considering that the ability of the model to detect apples regardless of image distortions is also crucial for real-world application of agronomic management, several data augmentation modes were employed to mitigate the severe performance degradation which usually caused by hardware-degraded or poor whether environment in actual application. Otherwise, in order to increase the networks capability of generalizing and reduce the probability of overfitting, the training set was corrupted with six image distortions, each spanning three levels of severity, as data augmentation (as shown in the last row of Figure 1). In addition, in order to adapt the trained algorithm to target recognition at low resolution more, the original images were cropped to the size of 4000×4000 pixels and further downscaled to 512×512 pixels. Finally, after filtering and cropping, 268 images were manually annotated using the labelme annotation tool. And without any additional labelling costs or architecture changes. 2813 images containing 5831 apples were totally generated for model training and 1000 images containing 2142 apples were totally generated for model evaluation. More detailed data set information is shown in Table 1. Otherwise, it should be noted that the training process drew support from transfer learning by migrating the pretrained model weights to RS-Net architecture before formal training, in which the pretrained weights were obtained by extracting the 1586 images containing 5851 apples from MS COCO (Lin et al., 2014) and then trained on RS-Net model. Through pre-training on these extracted images, model could accelerate convergence speed and get better performance.

Table 1 Image acquisition and data set division.

Specific	training images / instances	validation images / instances	total images / instances	pretraing set images / instances
Day	1581 / 2936	646 / 1153	2227 / 4089	- / -
Night	1232 / 2895	354 / 989	1586 / 3884	- / -
Total	2813 / 5831	1000 / 2142	3813 / 7973	2586 / 5851

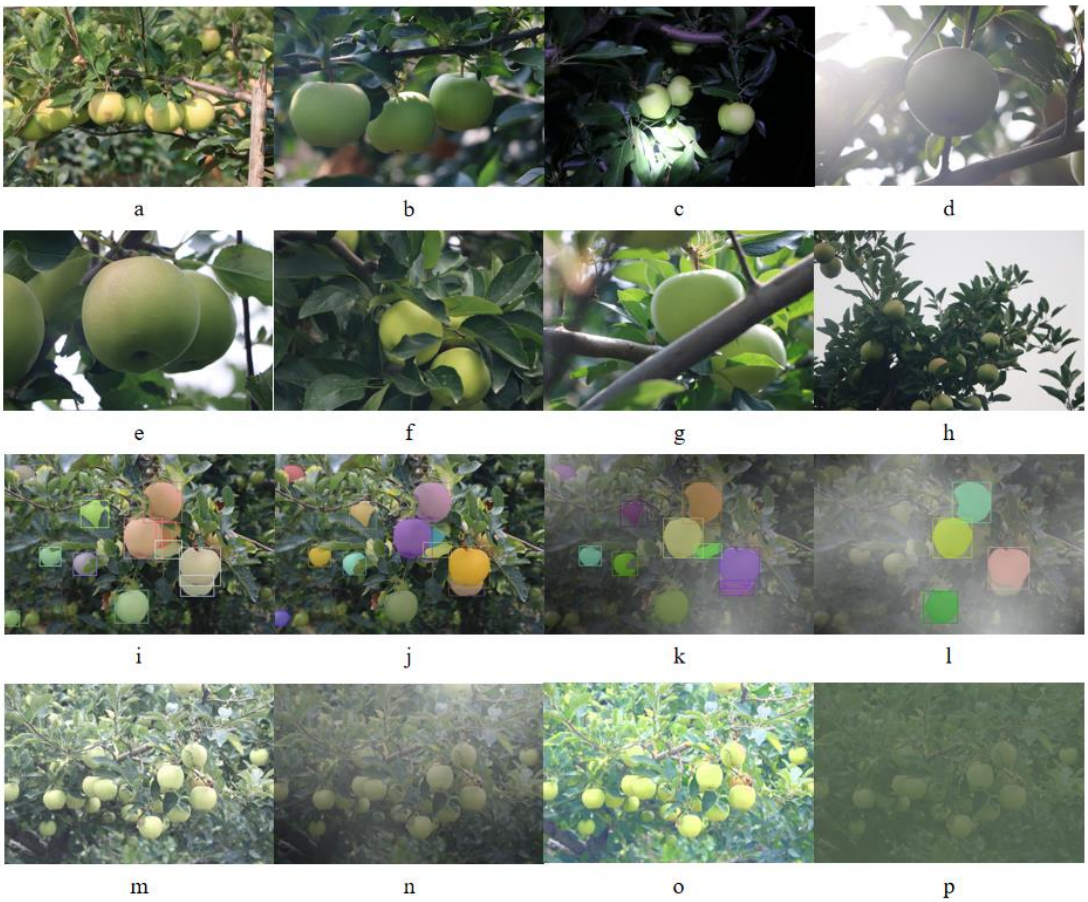

Fig. 1 a-d represent the images taken in different time intervals and different illumination angles; e-f show different types of occlusion(inter-fruits overlapped, leaves occlusion, branches occlusion and their combination);i is the ground truth corresponding to this image, and j-l are both segmented by Mask R-CNN which equipped with same weights and configurations. Apparently, as the fog gets thicker, the segmentation effect gets worse, which commonly appears in most segmentation methods. m-p show four of six corruption types (gaussian noise, impulse noise, brightness, fog, snow and contrast) with middle severity.

RS-Net

Mask R-CNN is a start-of-the-art instance segmentation algorithm which extends many previous excellent researches works (Shelhamer et al., 2017). This approach efficiently detects objects while simultaneously generating a high-quality segmentation mask for each instance in an image. In this paper, RS-Net is extended by original Mask R-CNN and make it more suitable for the segmentation of overlapped fruits in complex scenes. The overall pipeline of RS-Net is shown in Figure 2, It consists of three-part: (1) Feature Acquisition, (2) RoIs Generation, and (3) Results Prediction. Firstly, the pipeline of 'Feature Acquisition' consists of three steps: extraction, fusion and refining, which are respectively performed by ResNet, FPN and BFP (specifically in Figure 3). Then, based on the features generated by BFP, RPN produces abundant anchors on original images and outputs a set of object proposals that have been initially filtering. Finally, the mask is generated by FCN to indicate the detailed area where the apples are located.

The goal of RS-Net is to focus on the informative pixels but also suppress the noise by selectively aggregating the similar features of inconspicuous fruits, thus exploiting the potential of

the proposed model architectures for applying on vision system of agronomic management as much as possible. All components will be detailed in the following sections.

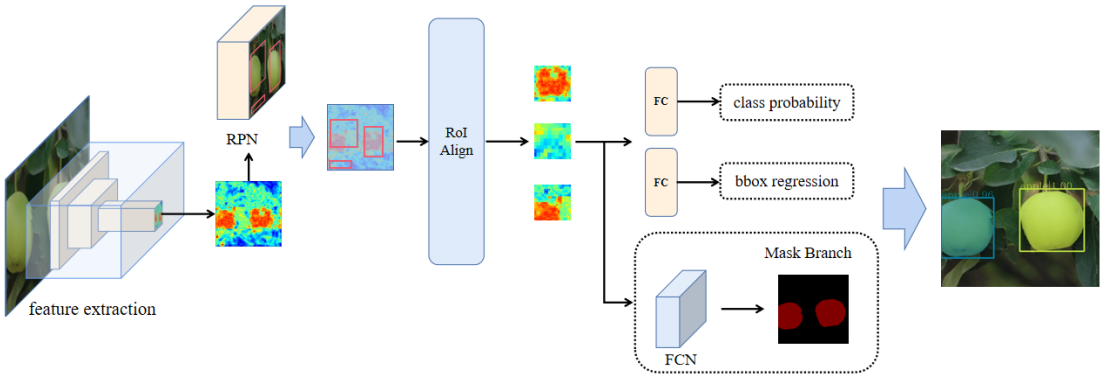

Fig. 2 Overview of the improved Mask R-CNN:an overall pipeline design for apple segmentation consists of three parts:(1) Feature Acquisition, (2) RoIs Generation, and (3) Results Prediction.

Feature Acquisition (ResNet+FPN+BFP)

The overall pipeline of 'Feature Acquisition' is shown in Figure 3. This section can be divided into three parts: extraction, fusion and refining, which are respectively performed by ResNet, FPN and BFP. Specifically, the combination of ResNet and FPN has been widely applied in many detection and segmentation architectures due to its excellent effect of feature representation, which also fits with the research goal of this paper. Generally, the depth of the network is crucial for learning the features with stronger representation ability, but with network depth increasing, it will bring about the problems such as gradient vanishing and explosion, which will lead to model degradation. In case the problems aforementioned, ResNet effectively solves this contradictory phenomenon by explicitly let shallower layers and deeper layers fit a residual mapping, thus improving the discriminative ability of the networks with deeper layers. According to the efficient feature extraction ability of ResNet, RS-Net could better mean and represent the image features on the basis of labeling information.

Generally, the output of last layer of ResNet has been provided sufficient semantic information, but also with the cost of missing detailed information related to object boundaries and resolution due to the consecutive down sampling operations (convolution and pooling), this will make the semantic information of smaller objects seriously diluted and finally cause the detection to fail. Considering that the design of vision system in agronomic management also needs to accurately recognize smaller area apples in an image due to the distance between sensors and objects, FPN is introduced to RS-Net architecture Typically, deep high-level features in backbones are with more semantic meanings while the shallow low-level features are more content descriptive. In other words, low level and high-level information is complementary in terms of semantic meanings and content details. Based on this point, FPN develops a top-down architecture with lateral connections for building high-level semantic feature maps at all scales, so as to improve the final accuracy on small area objects in this way. The details are shown in the left of Figure 3. Specifically, FPN uses the feature activations produced by each stage's last residual block of ResNet, and denotes the outputs

of these last residual blocks as $\{F_2, F_3, F_4, F_5\}$ for conv2, conv3, conv4 and conv5 stages. The set of feature maps integrated by FPN is called $\{A_2, A_3, A_4, A_5\}$, corresponding to $\{F_2, F_3, F_4, F_5\}$ that are respectively of the same spatial sizes.

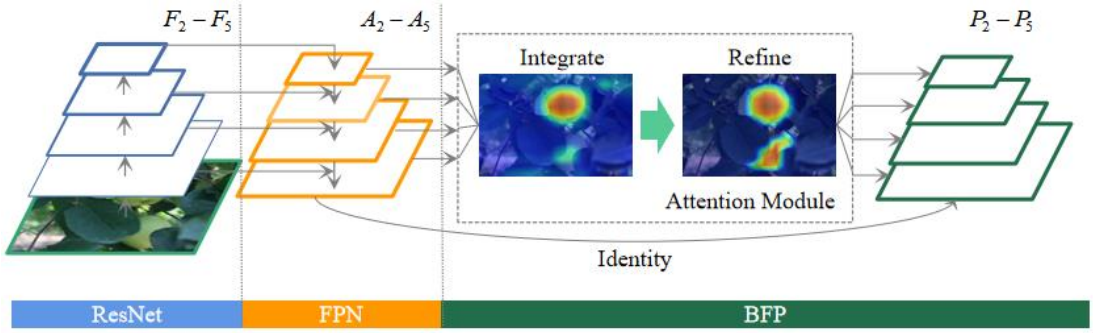

Fig. 3 Overall pipeline of 'Feature Acquisition' section. Images will be processed continuously as above to get the final finer feature maps (P_2-P_5) for the next steps. In this figure, feature maps are indicate by different color outlines, and thicker outlines denote semantically stronger features. Detailed pipeline of 'Attention Module' is illustrated in Figure 4.

Normally, the features via ResNet and FPN can be enough served as the basis for detection and segmentation, but considering two important factors, BFP module is introduced to the architecture for further refining the extracted features. The first point is that a large percentage of apples in collected images are inconspicuous or incomplete due to adverse factors such as lighting, occlusions, overlaps, etc., this will make the semantic information of inconspicuous fruits easily be disturbed by dominated salient object (e.g. leaves, branches) and diluted by consecutive down sampling operations. The second point is that some studies reveals that the best integrated features methods should possess balanced information from each resolution. But the sequential manner in FPN methods will make integrated features focus more on adjacent resolution but less on others. The semantic information contained in non-adjacent levels would be diluted once per fusion during the information flow. Therefore, in order to relieve the two aforementioned dilemmas simultaneously, BFP module is introduced the model architecture, which is illustrated in the right of Figure 3 and detailed in Figure 4.

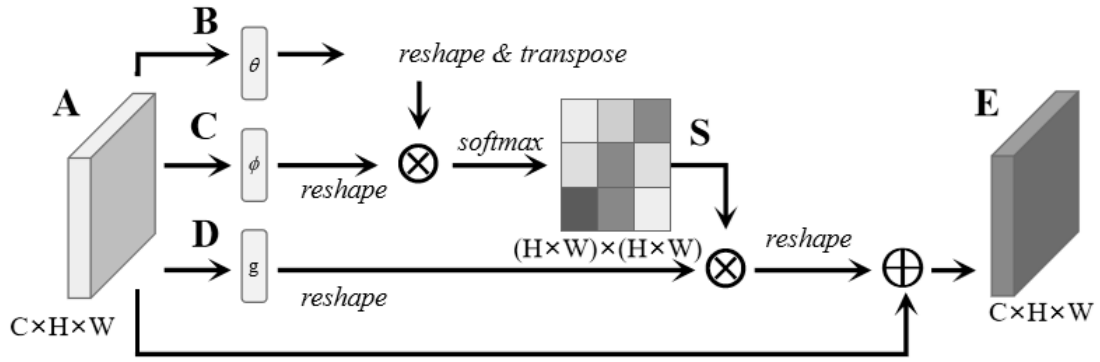

Fig. 4 Detailed description of attention module which illustrated in the BFP section of Figure 3

Features at level l and the number of features generated by FPN are respectively denoted as A_l and L. The indexes of involved smallest and biggest levels are denoted as $l_{\min}$ and $l_{\max}$. In Figure 3, A_2 has the biggest resolution. BFP first rescales the features $\{A_2, A_3, A_4, A_5\}$ to an

intermediate size A_4, with interpolation or adaptive max-pooling operation respectively. Finally, the balanced semantic features are obtained by simply averaging as:

$$A = \frac{1}{L} \sum_{l=l_{\min}}^{l_{\max}} A_l \tag{1}$$

Through this simple procedure as Eqn. (1) shown, each feature level contains equal information from others by resizing and averaging operations without any extra parameters. Next, the balanced semantic feature A will be further refined to get more discriminative by embedded Gaussian non-local operation, Firstly, a general formula for non-local operation is defined as Eqn. (2):

$$E_i = \frac{1}{C(x)} \sum_{\forall j} f(A_i, A_j) g(A_j) \tag{2}$$

Here $A \in R^{C \times H \times W}$ is the balanced semantic feature map and i denotes the position index whose similarity map will be computed, j denotes the index that enumerates all positions of A. f is the pairwise function to compute a scalar that represent the relationship between i and all j. E is the output signal of point i and with the same spatial size of A. The unary function g computes a representation of A at the position j, for simplicity, only consider g in the form of a linear embedding: $g(A_j) = D_j = W_g A_j$, where W_g is a weight matrix to be learned and implemented with 1×1 convolution. As for pairwise function, BFP employs embedding Gaussian function to compute the similarity.

Specifically, non-local operation first feeds A into 1×1 convolution layers(θ and ϕ) to generate two new feature maps B and C, respectively, where$\{B, C\} \in R^{C \times H \times W}$. Then it reshapes them to $R^{C \times N}$, where $N = H \times W$ is the number of pixels. After that BFP performs a matrix multiplication between the transpose of B and C, and applies a softmax layer to calculate the correlation intensity matrix between any two points $S \in R^{N \times N}$:

$$s_{ij} = \frac{\exp(B_i \bullet C_j)}{\sum_{j=1}^{N} \exp(B_i \bullet C_j)} \tag{3}$$

where s_{ij} measures the relationship between i^{th} position and j^{th} position. The more similar feature representations of the two positions contributes to greater correlation between them.

Meanwhile, non-local operation also feeds feature A into another convolution layer g to generate a new feature map $D \in R^{C \times H \times W}$ and reshapes it to $R^{C \times N}$. Then non-local operation performs a matrix multiplication between D and the transpose of S and reshapes the result to $R^{C \times H \times W}$. Finally, non-local operation performs a element-wise sum operation with the features A to obtain the final output $E \in R^{C \times H \times W}$ as follows:

$$E_i = \sum_{j=1}^{N} (s_{ij} \cdot D_j) + A_i \tag{4}$$

It can be inferred from Eqn. (4) that the resulting feature E at each position is a weighted sum of the features across all positions and original features. Therefore, it has a global contextual view and selectively aggregates contexts according to the correlation intensity matrix S. The similar semantic features achieve mutual gains, thus improving semantic similar information but also suppressing noises.

RoIs generation

For each feature map P_i in $\{P_2, P_3, P_4, P_5\}$ generated by last stage, it will be input into the RPN (Figure 5) to generate abundant anchors of different shapes, which are mapped to different apple shapes caused by overlaps and occlusions as possible. Then RPN initially filters the generated anchors given the probability of being a foreground. The architecture of RPN just consists of one 3*3 convolutional layer and followed by two 1*1 convolutional layers (for regression\classification, and denoted as reg\cls respectively), which is nearly cost-free given detection network computation. Concretely, 3*3 convlutional layer could be seen as a sliding-window to traverse all points at P_i, at each sliding-window center location, RPN simultaneously predicts multiple region anchors at original images. Considering that FPN has been adopted to alleviate scale variation, thus RPN only employs single area scale 8*8 with three aspect ratios (1:2, 1:1, 2:1) for each feature map level. For a convolutional feature map of a size W*H, there are 3*W*H anchors in total. Sequentially, cls is responsible for predicting the probability of each anchor being an foreground and reg is responsible for predicting a 4-D vector representing the 4 parameterized coordinates of the predicted bounding box for each anchor. Finally, Non-Maximum Suppression (NMS) is applied to filter out partial anchors based on the confidence scores predicted by cls and bbox offsets predicted by reg. The remaining anchors are the outputs of RPN, which are called 'proposals'. The embedding of RPN just makes the extra cost of two convolutional layers but act an important role in the while network structure.

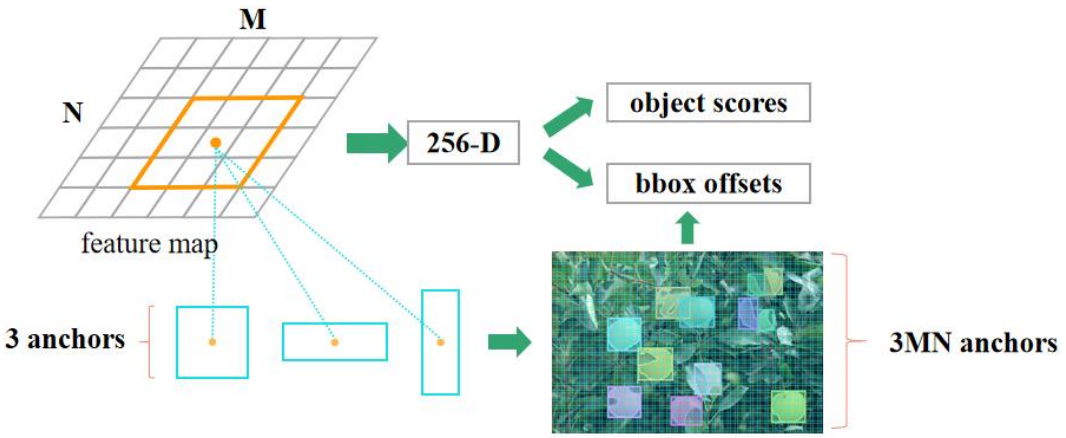

Fig. 5 Detailed description of RPN. 256-D represents a 256 dimensional vector after 3*3 convolution at each spatial location in feature map

Due to the proposals were generated at original images, model should map them into corresponding level to get features inside proposals, which are called Regions of Interest (RoIs). Since there are multiple feature maps owing to FPN, RoI Align layer needs to assign proposals of different scales to the certain pyramid level. Formally, the corresponding relationship between proposal (with width w and height h) to the level P_k of feature pyramid by:

$$k = \left\lfloor k_0 + \log_2\left(\sqrt{wh}/512\right)\right\rfloor \tag{5}$$

Here 512 is the uniform image size, and k_0 is the target level on which a proposal with $w \times h = 512 \times 512$ should be mapped into. Intuitively, Eqn. (5) means that if the area of proposal become bigger, it should be mapped into a coarser-resolution level. Next, RoIs are fed into RoI

Align layer improved from spatial pyramid pooling (SPP) for stretching them to same scale, which removed the harsh quantization of RoIPool and will play a key role in the next mask prediction.

Results Prediction

A RS-Net has three sibling output branches with different tasks for final predictions. The first outputs a probability distribution (per RoI) of being an apple. Although in the task of current work, only one category needs to be identified, the comprehensive evaluation metric AP which will be explained next needs the probability value to calculate precision and recall over each intersection of union (IoU) threshold, thus the model retains this branch for model evaluation and intuitive comparison with other methods. The second sibling layer outputs bounding-box regression offsets for adjusting proposals. Finally, the third branch employs Fully Convolution Network (FCN) at each RoI to achieving instance segmentation task. Specifically, this branch predicts a m*m mask from each RoI using an FCN without collapsing it into a vector representation that lacks spatial dimensions and make a pixel-wise prediction for each point in RoI through up- and down-sampling continuously. By combining the prediction results of three sibling branches, the final segmentation targets are obtained.

Implementation Details

Since a lot of hyper-parameters are needed in the implementation process, and the results are sensitive to the setting of these elements, thus these hyper-parameters are found for better segmentation performance by trial and error empirically.

In the training phase, the whole architecture can be trained end-to-end by stochastic gradient descent (SGD) and back propagation. Images first normalized with mean = [0.50, 0.42, 0.34]and std = [0.28, 0.27, 0.28] which are calculated from training dataset. ResNet50 is used as the main backbone to reduce the running time and publicly available. For each iteration, employ 2 images as a batch and BN (batch normalization) while updating weights. Initial learning rate, momentum and weight decay is set to 0. 0025, 0. 9 and 0. 0001 respectively, decrease it by 0. 1 after 8 and 11 epochs respectively if not specifically noted. Set base anchor scales and aspect ratios as 8 and [0. 5, 1, 2] while training RPN. As for loss function, the overall could be mainly divided into two parts: the losses of classification and offset regression from RPN section, and the multi-task losses from 'Results Prediction' section which includes classification branch, coordinate regression branch and mask segmentation branch. As shown in below:

$$\begin{aligned} L_{final} &= L_{RPN} + L_{Results-Prediction} \\ &= L_{cls1} + L_{reg1} + L_{cls2} + L_{reg2} + L_{mask} \end{aligned} \tag{6}$$

Here L_{final} denotes the final loss which will use for back propagation, L_{RPN} consists of L_{cls1}, L_{reg1} and $L_{Results-Prediction}$ consists of L_{cls2}, L_{reg2} and L_{mask} represent losses from RPN section and Results Prediction section respectively. Specifically, model employs cross entropy loss function for L_{cls1}, L_{cls2} and L_{mask}, and L1 loss function for L_{reg1} and L_{reg2}. For each feature level generated by BFP, 256 anchors are randomly sampled as a mini-batch for computing L_{RPN}, where sampled negative anchors and positive anchors with a 1:1 ratio. Replace the batch with negative ones if there are fewer than 128 positive samples in the original image.

Experiments

Evaluation metric

In order to evaluate the detection performance more comprehensively and strictly, AP (average precision) is employed as main evaluation metric which averages the precision values calculated over IoUs from 0.5 to 0.95 with an interval of 0.05. Firstly, define I as a set of equally spaced IoUs thresholds levels [0.5, 0.55, …, 0.95]. For each threshold i in I , if the IoU between predicted bbox and the matched ground truth exceeds i , this example is defined as true positive (*TP*) example, else, as false positive (*FP*), and the ground truth which are not detected successfully by detector is defined as false negative (*FN*). Then, at most the top 100 predicted bboxes given confidence scores are selected and then used to calculate the precision (*P*) and recall (*R*) (Eq. (7)) pair corresponding to sorted confidence thresholds in turn, thus the precision/recall pairs over a specific IoU threshold and multiple confidence thresholds are calculated.

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN} \tag{7}$$

AP over a specific IoU threshold i could be seen as the approximate area under the precision/recall curve (AUC), and is defined as the mean precision at a set of 101 equally spaced recall levels R : [0, 0. 01, . . . , 1]:

$$AP^{IoU=i} = \frac{1}{101} \sum_{r \in R} p_{interp}(r) \tag{8}$$

The precision at each recall level *r* is interpolated by taking the maximum precision measured from which the corresponding recall exceeds r:

$$p_{interp}(r) = \max_{\tilde{r}:\tilde{r} \geq r} p(\tilde{r}) \tag{9}$$

where $p(\tilde{r})$ is the measured precision at recall $\tilde{r}$. Similarly, all $AP^{IoU=i}(i \in I)$ could get by following the above steps and the final evaluation metric AP could be formulated as:

$$AP = \frac{1}{10} \sum_{i \in I} AP^{IoU=i} \tag{10}$$

The factor "10" corresponds to the number of the IoUs thresholds tested in set *I* . Intuitively, *AP* evaluates the result over different IoU thresholds, confidence scores, precisions and recalls, thus can measure RS-Net accurately and comprehensively. Both box *AP* and mask *AP* are evaluated. In addition, *AR* (average recall) is also used as an evaluation metric, which is obtained by taking the average value of $AR^{IoU=i}$s over 10 IoU thresholds tested given the top 100 predicted bboxes at most. Since the task of the model only needs to identify one category, $AR^{IoU=i}$ under a specific threshold is equal to *R* in Eq. (7). More information about evaluation metrics please refer to MS COCO for detailed explanation.

Model training

Totally, 2813 images containing 5831 apples are used for training process. RS-Net is trained with 12 epochs and a total of 16884 iterations (2 images/iteration). In addition, despite dataset is extended over different corruptions, due to that there is only one category, which makes the training

process easier to overfitting. To eliminate this hidden trouble and accelerate network convergence, RS-Net is pretrained over 1586 images which extracted from MS COCO dateset without extra annotation works and then loads the pretrained weights into model architecture as initialization parameters for formal training. Intuitively, the loss value curve changes with iterations on the above two situations are illustrated in Figure 6.

Obviously, thicker curve begins at a remarkably smaller value than thinner one and the loss value is about 0.1 smaller when the end of two curves tend to be stable. It can be inferred from this figure that the formal training of model can get benefit from 1586 images used for pre-training, which makes the model learn more distinguishing features and less risk of over-fitting. Comparing the obvious gap between two results, the pre-training way is adopted to carry out the following processes for better performance.

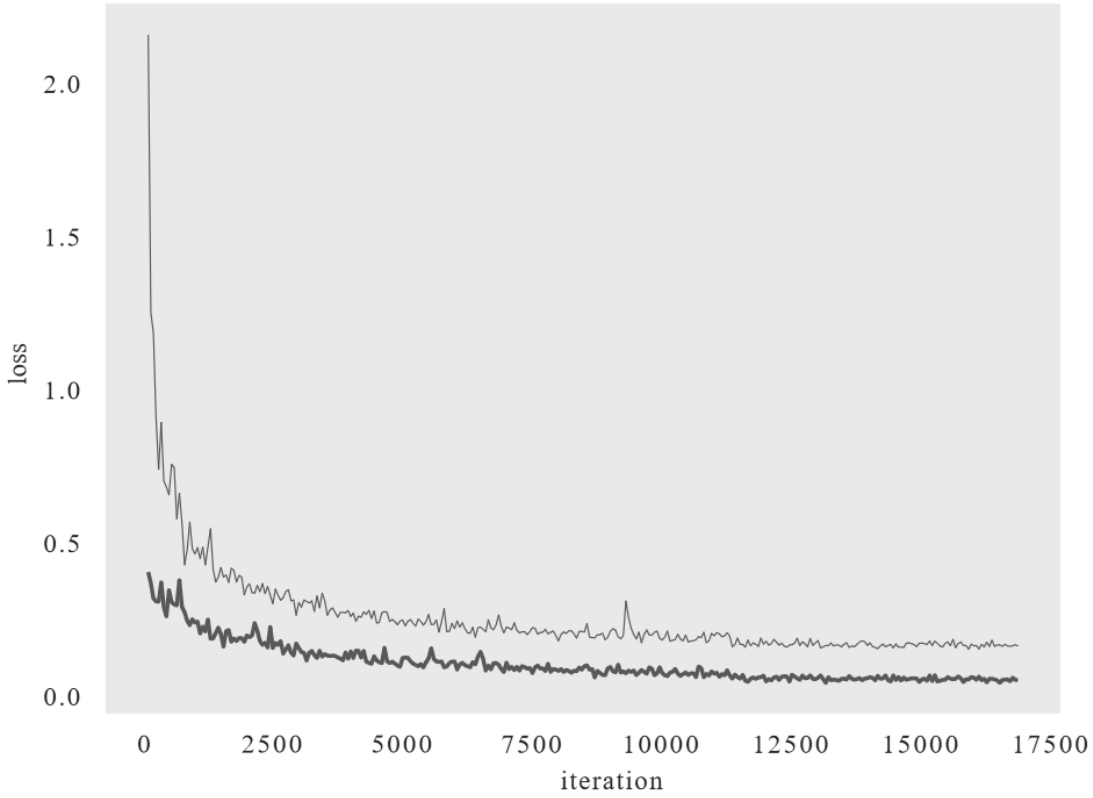

Fig. 6 Loss value curve changes with iterations. Thicker curve represents training process equipped with pretained weights and thinner curve represents no pre-training.

Ablation Experiments

For fair comparisons in ablation experiments and validate the effect of attention module, experiment employs original Mask R-CNN built on MMDetection v2.0 (Chen et al., 2018) as baseline and both same hyper-parameters as RS-Net except section of BFP. Since Mask R-CNN and RS-Net both have a relatively good segmentation effect, thus experiment directly employs IoU=0. 90 as strict threshold for defining a bbox as *TP* or *FP* to measure the high-quality effect gap between the baseline and RS-Net. Table 2 lists the specific comparison results of two methods.

Table 2 Specific comparison results of two methods

Method	AP_{90}^{box} /%	AR_{90}^{box} /%	AP_{90}^{mask} /%	AR_{90}^{mask} /%
Baseline	78.8	80.0	80.8	80.7

RS-Net	82.6	82.7	81.1	84.6

As shown in Table 1, except for average precision, RS-Net can also enhance the average recall of predicted bboxes. The embedding of attention module brings 3.7 points higher AR_{90}^{box} and 2.9 points higher AR_{90}^{mask} compared with baseline method. This phenomenon is due to that BFP could make the similar semantic features achieve mutual gains across inconspicuous and salient apples, the heavily inconspicuous apples caused by overlapped, occlusion and illumination could get help from salient apples, thus the proportion of boxes judged as *TP* in the whole ground truth will rise and *AR* metrics will be significantly improved. Several images containing heavily incomplete apples are visualized for intuitively feeling the gap between the two in Figure 7.

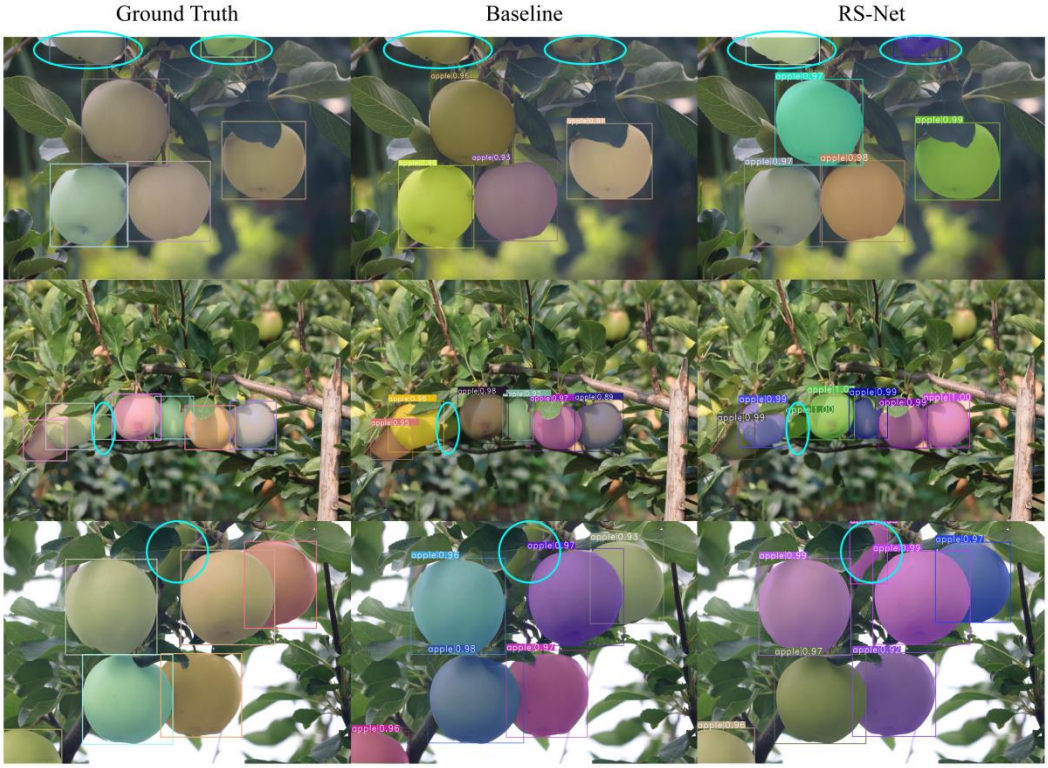

Fig. 7 Visualization of comparison results over two methods. Ellipses represent apples that were labeled as ground truth, and the baseline method did not detect it successfully but RS-Net did. Circles represent apple that were not labeled as ground truth due to severe occlusion, but RS-Net still detected it.

As shown in the above figure, two methods both have good segmentation effect when detecting conspicuous apples, However, due to the attention mechanism employed in RS-Net, severely occluded apples can also be well segmented by drawing information from salient parts, which even includes the severely occluded apples that are not labeled as ground truth. This is also the reason why RS-Net get higher metric values. Therefore, RS-Net is better in segmenting overlapped apples in the same color background and more suitable for deploying on vision system of harvesting robot.

Comparison with state-of-the-arts methods

For further validation of the improved Mask R-CNN, experiments compare the proposed model with the state-of-the-art detection and instance segmentation methods with identical experimental configuration. It should be noted that all experiments reported in current work are tested on the same environment equipped with Tesla V100 GPU, CUDA V10.0, and Pytorch 1.4 for studies.

Detection Effect

Since the main body of RS-Net is extended on the detector architecture by adding a mask branch, and the mask segmentation is operated based on the predicted boxes, in other words, the detection effect of the model directly affects the segmentation effect, thus experiments first compare the detection effect of RS-Net with the start-of-the-art detectors. As for evaluation metrics, in addition to using the box Average Precision (AP^{box}) metric which averages *AP*s across IoU thresholds from 0.5 to 0.95 with an interval of 0.05, AP^{box}_{50} and AP^{box}_{90} (AP at different IoU thresholds) are reported as loose and strict boundaries, respectively. The specific comparison results are shown in Table 3.

Table 3 Comparison with state-of-the-art detection methods on validation dateset

Method	Backbone	AP^{box}/%	AP^{box}_{50} /%	AP^{box}_{90}/%
SSD512	VGG16	78.2	88.5	70.9
Faster R-CNN	R-50-FPN	83.4	89.0	77.5
RetinaNet	R-50-FPN	80.6	88.4	73.6
Mask R-CNN	R-50-FPN	84.5	88.4	78.8
RS-Net	R-50-FPN-BFP	**85.6**	**90.0**	**82.6**

Intuitively, the detection effect of original Mask R-CNN outperforms the other advanced detectors with the same extraction capability of backbone while detecting on test set. It achieves 7.4%, 1.1% and 3.9% AP^{box} gains compared with SSD (Liu et al., 2016a), Faster R-CNN, RetinaNet (Lin et al., 2020) respectively. By adding attention module to naive Mask R-CNN, the improved Mask R-CNN obtains further performance which brings 1.1%, 1.6% and 3.8% gains in terms of AP^{box}, AP^{box}_{50} and AP^{box}_{90}. Comprehensively, from the above analysis, RS-Net achieves better detection effect, which could be more suitable and robust for deploying on vision system of apple harvesting robots.

Segmentation Effect

Due to the aim of this paper is to explore the ability of the model to segment overlapped fruits in the same color background, thus in-depth comparative experiments with start-of-the-art instance segmentation methods are carried out and experimental results of them are analyzed to validate the effectiveness of RS-Net. The specific comparison results are shown in Table 4.

Table 4 Comparison with state-of-the-art instance segmentation methods on validation dateset.

Method	AP^{box}/%	AP^{mask}/%	AP^{mask}_{50} /%	AP^{mask}_{90}/%
YOLACT	67.4	75.6	89.1	69.8
YOLACT++	78.0	78.8	89.1	76.3
RetinaMask	82.8	83.6	88.8	82.2
RS-Net	**85.6**	**86.2**	**90.0**	**84.1**

All models employed ResNet50 as feature extractor for fair comparison. In contrast, RS-Net achieves the best results in terms of both box *AP* and mask *AP* metrics. In particular, compared to RetinaMask (Fu CY et al., 2019) which has similar architecture (detector + mask branch), RS-Net achieves 2.8% AP^{box} gain and 2.6% AP^{mask} gain respectively. Otherwise, it should be noted that the gap between AP^{mask}, AP^{mask}_{50} and AP^{mask}_{90} is smaller than YOLACT (Bolya et al., 2019), YOLACT++ (Bolya et al., 2020) and RetinaMask, which means that the most masks segmented by

RS-Net are concentrated in high quality area (higher IoU with ground truth). In order to more intuitively feel the effect of RS-Net, several representative images which containing different numbers of apples are selected and used different methods to segment. The visualization results are shown in Figure 8. Obviously, the effect of the proposed method is much better than other methods in terms of both recognition accuracy and segmentation effect. In addition, since RS-Net introduces attention module into architecture, most heavily overlapped apples are also well segmented, including some are not even labeled as ground truth.

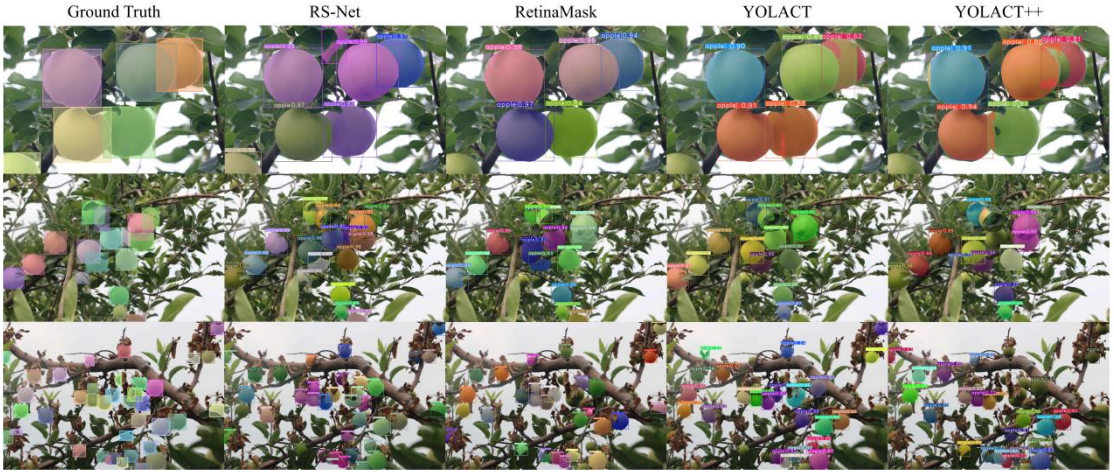

Fig. 8 Visualization results of different test images segmented by RS-Net, RetinaMask, YOLACT and YOLACT++ respectively.

Conclusion

In order to effectively detect overlapped fruits in natural environment, RS-Net architecture which extends Mask R-CNN by adding an embedding Gaussian attention module is proposed, thus make the similar semantic features achieve mutual gains and reduce the impact of adverse factors such as occlusion, illumination, overlapped, etc. The experimental result shows that the proposed RS-Net outperforms other start-of-the-art deep learning-based methods when applying on vision system of harvesting robot, and achieve both higher accuracy and stronger robustness, which could be more suitable for operating in real-world scene for harvesting robot's vision system.

Although RS-Net has achieved relatively ideal experimental results, there are still some aspects and rooms that need to be improved continuously in architecture. For example, the average segmentation time of each image with size 512×512 on GPU is 65.79 ms, while the fastest model in experiments (YOLACT++) only needs 20.43 ms. The shortest inference times of other researches which also reported on 512*512 resolutions in fruit detection and segmentation field need 15ms (Koirala et al., 2019b) and 390ms (Li et al., 2021) respectively. In contrast, the inference time of RS-Net is in a lower middle position. Though this has been able to meet the real-time needs of practical deployment, it is still a little longer than other methods in terms of time-consuming. This phenomenon is suspected to be caused by two reasons: 1) The Faster R-CNN which Mask R-CNN extends is a two-stage architecture for better accuracy, which will inevitably lead to bigger consumption of computation and time than other one-stage methods. 2) The RS-Net is anchor-based for achieving a higher recall rate, which will require the model to densely place anchor boxes on the original images, it also leads to more time-consuming. Based on this defect, extending mask branch

to other one-stage or anchor-based detection methods is considered to strike the better trade-off between speed and accuracy simultaneously in the future works.

Acknowledgments

This work is supported by Natural Science Foundation of Shandong Province in China (No.: ZR2020MF076) Focus on Research and Development Plan in Shandong Province (No.: 2019GNC106115); National Nature Science Foundation of China (No.: 62072289); Shandong Province Higher Educational Science and Technology Program (No.: J18KA308); Taishan Scholar Program of Shandong Province of China (No.: TSHW201502038).

Declarations

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Reference

Aggelopoulou, A. D., Bochtis, D., Fountas, S., Swain, K. C., Gemtos, T. A., & Nanos, G. D. (2011). Yield prediction in apple orchards based on image processing. *Precision Agriculture*, 12(3), 448-456. https://doi. org/10.1007/s11119-010-9187-0

Cheein, F. A. A., & Carelli, R. (2013). Agricultural robotics: Unmanned robotic service units in agricultural tasks. *IEEE industrial electronics magazine*, 7(3), 48-58. https://doi. org/10.1109/MIE.2013.2252957

Bac, C. W., van Henten, E. J., Hemming, J., & Edan, Y. (2014). Harvesting robots for high‐value crops: State‐of‐the‐art review and challenges ahead. *Journal of Field Robotics*, 31(6), 888-911. https://doi.org/10.1002/rob.21525

Bargoti, S., & Underwood, J. (2017a). Deep fruit detection in orchards. In IEEE International Conference on Robotics and Automation (ICRA), pp. 3626-3633. https://doi. org/10.1109/ICRA.2017.7989417

Bargoti, S., & Underwood, J. P. (2017b). Image segmentation for fruit detection and yield estimation in apple orchards. *Journal of Field Robotics,* 34(6), 1039-1060. https://doi. org/10.1002/rob.21699

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2019). Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157-9166. https://doi. org/10.1109 /ICCV.2019.00925

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. J. (2020). Yolact++: Better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. https://doi. org/10.1109/TPAMI. 2020.3014297

Chen, K., Pang, J., Wang, J., et al. mmdetection. https://github.com/open-mmlab/ mmdetection, 2018.

Chen, L. C., Yang, Y., Wang, J., Xu, W., & Yuille, A. L. (2016). Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640-3649. https://doi. org/10.1109/CVPR.2016.396

Chen, S. W., Shivakumar, S. S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C., & Kumar, V. (2017). Counting apples and oranges with deep learning: A data-driven approach. *IEEE Robotics and Automation Letters*, 2(2), 781-788. https://doi. org/10.1109/LRA.2017.2651944

Dodge, S., & Karam, L. (2016, June). Understanding how image quality affects deep neural networks. In 2016 eighth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6. https://doi. org/10.1109/QoMEX.2016.7498955

Fu, C. Y., Shvets, M., & Berg, A. C. (2019a). RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. arXiv preprint arXiv:1901.03353.

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019b). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146-3154. https://doi. org/10.1109/CVPR.2019.00326

Fu, L., Majeed, Y., Zhang, X., Karkee, M., & Zhang, Q. (2020a). Faster R–CNN–based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. *Biosystems Engineering*, 197, 245-256. https://doi. org/ 10.1016/j.biosystemseng.2020.07.007

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., Cao, Q., Tian, Y., Zhu, Y., Cao, W., & Liu, X. (2020b). Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle. *Remote Sensing*, 12(3), 508. https://doi.org/10.3390/rs12030508

Gené-Mola, J., Vilaplana, V., Rosell-Polo, J. R., Morros, J. R., Ruiz-Hidalgo, J., & Gregorio, E. (2019). Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. *Computers and electronics in agriculture*, 162, 689-698. https://doi.org/10.1016/j.compag.2019.05.016

Gongal, A., Amatya, S., Karkee, M., Zhang, Q., & Lewis, K. (2015). Sensors and systems for fruit detection and localization: A review. *Computers and Electronics in Agriculture*, 116, 8-19. https://doi.org/10.1016/j.compag.2015.05.021

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, pp. 2961-2969. https://doi. org/10.1109/ICCV.2017.322

He, W., Zhang, X. Y., Yin, F., Luo, Z., Ogier, J. M., & Liu, C. L. (2020). Realtime multi-scale scene text detection with scale-based region proposal network. *Pattern Recognition*, 98, 107026. https://doi. org/10.1016/j.patcog.2019.107026

Hung, C., Underwood, J., Nieto, J., & Sukkarieh, S. (2015). A feature learning based approach for automated fruit yield estimation. In Field and Service Robotics, pp. 485-498. https://doi.org/10.1007/978-3-319-07488-7_33

Ji, W., Zhao, D., Cheng, F., Xu, B., Zhang, Y., & Wang, J. (2012). Automatic recognition vision system guided for apple harvesting robot. *Computers & Electrical Engineering*, 38(5), 1186-1195. https://doi. org/10.1016/j.compeleceng.2011.11.005

Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., & Zheng, Y. (2020a). Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Computers and Electronics in Agriculture*, 172, 105380. https://doi. org/10.1016/j.compag.2020.105380

Jia, W., Zhang, Y., Lian, J., Zheng, Y., Zhao, D., & Li, C. (2020b). Apple harvesting robot under information technology: A review. *International Journal of Advanced Robotic Systems*, 17(3), 25310. https://doi. org/10.1177/1729881420925310

Jia, W., Zhao, D., Liu, X., Tang, S., Ruan, C., & Ji, W. (2015). Apple recognition based on K-means and GA-RBF-LMS neural network applicated in harvesting robot. *Transactions of the Chinese Society of Agricultural Engineering*, 31(18), 175-183. (in Chinese)

Kapach, K., Barnea, E., Mairon, R., Edan, Y., & Ben-Shahar, O. (2012). Computer vision for fruit harvesting robots–state of the art and challenges ahead. *International Journal of Computational Vision and Robotics*, 3(1-2), 4-34. https://doi.org/10.1504/IJCVR.2012.046419

Kim, S. W., Kook, H. K., Sun, J. Y., Kang, M. C., & Ko, S. J. (2018). Parallel feature pyramid network for object detection. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 234-250. https://doi.

org/10.1007/978-3-030-01228-1_15

Koirala, A., Walsh, K. B., Wang, Z., & McCarthy, C. (2019a). Deep learning–Method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture*, 162, 219-234. https://doi.org/10.1016/j.compag.2019.04.017

Koirala, A., Walsh, K. B., Wang, Z., & McCarthy, C. (2019b). Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture*, 20(6), 1107-1135. https://doi.org/10.1007/ s11119-019-09642-0

Kurtulmus, F., Lee, W. S., & Vardar, A. (2011). Green citrus detection using 'eigenfruit', color and circular Gabor texture features under natural outdoor conditions. *Computers and Electronics in Agriculture*, 78(2), 140-149. https://doi. org/10.1016/j.compag.2011.07.001

Larese, M. G., Namías, R., Craviotto, R. M., Arango, M. R., Gallo, C., & Granitto, P. M. (2014). Automatic classification of legumes using leaf vein image features. *Pattern Recognition*, 47(1), 158-168. https://doi. org/10.1016/j.patcog.2013.06.012

Li, Q., Jia, W., Sun, M., Hou, S., & Zheng, Y. (2021). A novel green apple segmentation algorithm based on ensemble U-Net under complex orchard environment. *Computers and Electronics in Agriculture*, 180, 105900. https://doi.org/10.1016/j.compag.2020.105900

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollar, P. (2020). Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 42(2): 318-327. https://doi. org/10.1109/TPAMI.2018.2858826

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In European Conference on Computer Vision, pp. 740-755. https://doi. org/10.1007/978-3-319-10602-1_48

Linker, R. (2018). Machine learning based analysis of night-time images for yield prediction in apple orchard. *Biosystems Engineering*, 167, 114-125. https://doi.org/10.1016/j.biosystemseng.2018.01.003

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016a). Ssd: Single shot multibox detector. In European Conference on Computer Vision, pp. 21-37. https://doi. org/10.1007/978-3-319-46448-0_2

Liu, X., Jia, W., Ruan, C., Zhao, D., Gu, Y., & Chen, W. (2018). The recognition of apple fruits in plastic bags based on block classification. *Precision Agriculture*, 19(4), 735-749. https://doi. org/10.1007 /s11119-017-9553-2

Liu, X., Zhao, D., Jia, W., Ruan, C., Tang, S., & Shen, T. (2016b). A method of segmenting apples at night based on color and position information. *Computers and Electronics in Agriculture*, 122, 118-123. https://doi. org/ 10.1016/j.compag.2016.01.023

Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., Bethge, E., & Brendel, W. (2019). Benchmarking robustness in object detection: Autonomous driving when winter is coming. arXiv preprint arXiv:1907.07484.

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., & Lin, D. (2019). Libra r-cnn: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 821-830. https://doi. org/10.1109/CVPR.2019.00091

Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: fruit counting based on deep simulated learning. *Sensors*, 17(4), 905. https://doi.org/10.3390/s17040905

Rakun, J., Stajnko, D., & Zazula, D. (2011). Detecting fruits in natural scenes by using spatial-frequency based texture analysis and multiview geometry. *Computers and Electronics in Agriculture*, 76(1), 80-88. https://doi. org/ 10.1016/j.compag.2011.01.007

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137-1149. https://doi. org/10.1109/TPAMI.2016.2577031

Zhou, R., Damerow, L., Sun, Y., & Blanke, M. M. (2012). Using colour features of cv.'Gala'apple fruits in an orchard in image processing to predict yield. *Precision Agriculture*, 13(5), 568-580. https://doi.org/10.1007/s11119-012-9269-2

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., & McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors*, 16(8), 1222. https://doi.org/10.3390/s16081222

Siegel, K. R., Ali, M. K., Srinivasiah, A., Nugent, R. A., & Narayan, K. V. (2014). Do we produce enough fruits and vegetables to meet global health need?. *PloS one*, 9(8), e104059. https://doi.org/10.1371/journal.pone.0104059

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4): 640-651. https://doi. org/10.1109/ TPAMI.2016.2572683

Targ, S., Almeida, D., & Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

Tian, Y., Duan, H., Luo, R., Zhang, Y., Jia, W., Lian, J., Zheng, Y., Ruan, C., & Li, C. (2019). Fast recognition and location of target fruit based on depth information. *IEEE Access*, 7, 170553-170563. https://doi. org/10.1109/ACCESS.2019.2955566

Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E., & Liang, Z. (2019). Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers and Electronics in Agriculture*, 157, 417-426. https://doi. org/ 10.1016/j.compag.2019.01.012

Underwood, J. P., Hung, C., Whelan, B., & Sukkarieh, S. (2016). Mapping almond orchard canopy volume, flowers, fruit and yield using lidar and vision sensors. *Computers and Electronics in Agriculture*, 130, 83-96. https://doi.org/10.1016/j.compag.2016.09.014

Vasconez, J. P., Delpiano, J., Vougioukas, S., & Cheein, F. A. (2020). Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation. *Computers and Electronics in Agriculture*, 173, 105348. https://doi.org/10.1016/j.compag.2020.105348

Wang, Q., Nuske, S., Bergerman, M., & Singh, S. (2013). Automated crop yield estimation for apple orchards. In Experimental robotics, pp. 745-758. https://doi.org/10.1007/ 978-3-319-00065-7_50

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794-7803. https://doi.org/10.1109 /CVPR.2018.00813

Wei X S, Xie C W, Wu J, et al. Mask-CNN: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 2018, 76: 704-714. https://doi. org/10.1016/j.patcog.2017.10.002

Yu, Y., Zhang, K., Yang, L., & Zhang, D. (2019). Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. *Computers and Electronics in Agriculture*, 163, 104846. https://doi. org/ 10.1016/j.compag.2019.06.001

Zhang, J., Huang, Y., Pu, R., Gonzalez-Moreno, P., Yuan, L., Wu, K., & Huang, W. (2019). Monitoring plant diseases and pests through remote sensing technology: A review. *Computers and Electronics in Agriculture*, 165, 104943. https://doi. org/10.1016/j.compag.2019.104943

Zhang, Z., Heinemann, P. H., Liu, J., Baugher, T. A., & Schupp, J. R. (2016). The development of mechanical apple harvesting technology: A review. *Transactions of the ASABE*, 59(5), 1165-1180. https://doi. org/10.13031/trans.59.11737

Zhong, Z., Sun, L., & Huo, Q. (2019). Improved localization accuracy by LocNet for Faster R-CNN based text detection in natural scene images. *Pattern Recognition*, 96, 106986. https://doi.org/10.1016/j.patcog.2019.106986

Authors and Affiliations

Weikuan Jia[1, 2], Zhonghua Zhang[1], Wenjiang Shao[1], Ze Ji[3], Sujuan Hou[1, 4]

[1] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China

[2] Key Laboratory of Facility Agriculture Measurement and Control Technology and Equipment of Machinery Industry, Zhenjiang 212013, China

[3] School of Engineering, Cardiff University, Cardiff, CF24 3AA, United Kingdom

[4] Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Shandong Normal University, Jinan 250358, China