Quantitative Biology 2022, 10(3): 239–252
https://doi.org/10.15302/J-QB-021-0272

RESEARCH ARTICLE

Lesion region segmentation via weakly supervised learning

Ran Yi[1], Rui Zeng[1], Yang Weng[1], Minjing Yu[2,*], Yu-Kun Lai[3], Yong-Jin Liu[1,*]

[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[3] School of Computer Science and Informatics, Cardiff University, Cardiff, CF10 3AT, United Kingdom
* Correspondence: minjingyu@tju.edu.cn; liuyongjin@tsinghua.edu.cn

Received January 31, 2021; Revised March 13, 2021; Accepted April 7, 2021

***Background*: Image-based automatic diagnosis of field diseases can help increase crop yields and is of great importance. However, crop lesion regions tend to be scattered and of varying sizes, this along with substantial intra-class variation and small inter-class variation makes segmentation difficult.**
***Methods*: We propose a novel end-to-end system that only requires weak supervision of image-level labels for lesion region segmentation. First, a two-branch network is designed for joint disease classification and seed region generation. The generated seed regions are then used as input to the next segmentation stage where we design to use an encoder-decoder network. Different from previous works that use an encoder in the segmentation network, the encoder-decoder network is critical for our system to successfully segment images with small and scattered regions, which is the major challenge in image-based diagnosis of field diseases. We further propose a novel weakly supervised training strategy for the encoder-decoder semantic segmentation network, making use of the extracted seed regions.**
***Results*: Experimental results show that our system achieves better lesion region segmentation results than state of the arts. In addition to crop images, our method is also applicable to general scattered object segmentation. We demonstrate this by extending our framework to work on the PASCAL VOC dataset, which achieves comparable performance with the state-of-the-art DSRG (deep seeded region growing) method.**
***Conclusion*: Our method not only outperforms state-of-the-art semantic segmentation methods by a large margin for the lesion segmentation task, but also shows its capability to perform well on more general tasks.**

Keywords: weakly supervised learning; lesion segmentation; disease detection; semantic segmentation; agriculture

Author summary: Crop diseases seriously affect the quantity and quality of crop yields, causing huge economic losses and posing a serious threat to global food security. However, overuse of chemicals in traditional agriculture may be harmful to humans and livestock. Therefore, early diagnosis of crop diseases, which helps to avoid the heavy use of chemicals and provides feasible solutions, is much desired. We proposed a system to segment images with small and scattered regions. Experimental results show that our method not only outperforms state-of-the-art segmentation methods for the lesion segmentation task, but also shows its capability to perform well on more general tasks.

INTRODUCTION

Crop diseases seriously affect the quantity and quality of crop yields, causing huge economic losses and posing a serious threat to global food security [1]. Traditional agriculture relies on heavy use of chemicals such as fungicides and insecticides to control crop diseases. However, overuse of chemicals may result in the problem of pesticide residues, toxic and harmful to humans and livestock, and even has a negative effect on agro-ecosystems [2]. Therefore, early diagnosis of crop diseases, which helps to avoid the heavy use of chemicals and provides feasible solutions, is much desired. In traditional agriculture, the identification of

crop diseases and the assessment of the infection severity are based on naked eye observation. These manual methods are not only time-consuming and laborious, but also inaccurate in estimated results. Accordingly, specially designed computer techniques that can achieve automatic diagnosis of crop diseases are very important.

From the perspective of plant pathology, diseases on leaves are abnormal manifestations after infection by bacteria, viruses or fungi. Often, diseased leaves show abnormal colors (*e.g.*, chlorosis, yellowing or whitening), and local or large area tissue necrosis and spots. Therefore, the images of impaired leaves are widely used as the subjects in automatic diagnosis of crop disease. Most of the current research efforts in this topic can be divided into two aspects: (1) The classification of disease types; (2) The judgement of disease severity, which can be evaluated by proportion of the diseased area on plants. Due to the excellent performance of machine learning technologies, the crop disease classification [3–6] and detection [7–9] achieved great progress.

Image-based classification and segmentation of crop diseases have been widely studied in computer vision. For crop disease classification, many works make use of deep neural networks, either from scratch or fine-tuning classic network models. Aravind *et al.* [3] applied transfer learning on a pretrained AlexNet [10] to classify grape diseases extracted from the PlantVillage dataset. Mohanty *et al*. [7] trained AlexNet [10] and GoogLeNet [11] to identify 26 diseases of 14 crop species and achieved good results. Dechant *et al*. [4] proposed a convolutional-neural-network (CNN) based three-stage process to analyze images for determining whether they contain infected leaves. In this work, diseased images need to be annotated by lines along the main axis of the lesion regions. Pound *et al*. [5] constructed four stacked hourglass networks for positioning the regions of wheat spikes and spikelets, and classifying the wheat type at the same time. Some works used more comprehensive, systematic approaches to perform classification. Krause *et al*. [12] proposed a pipeline for classification of plant species. It first uses a multi-scale approach to segment the region of interest. Then it trains a deep CNN through the multi-scale patches to determine the plant's category. Kumar *et al*. [13] proposed a mobile system called Leafsnap. The system classifies leaf types by extracting the contour curvature features of the leaves and comparing the features with those in the database including 184 trees. Fuentes *et al*. [14] proposed a deep-learning-based real-time system for identifying and locating tomato plant diseases and insect pests. The authors compared different network structures, including Faster R-CNN, SSD and R-FCN, combined with different feature extraction networks (such as VGG [15] and ResNet [16]). The method requires the lesion regions be annotated with bounding boxes and classes. For crop disease segmentation, Chen *et al*. [17] proposed a functional model that completes the leaf segmentation task and accurately depicts the leaf edge. Phadikar *et al*. [18] used the Fermi energy-based segmentation method to segment the lesion region of the rice disease from its background. Johannes *et al*. [19] used a random-forest based classifier trained with color and texture descriptors and a Bayesian segmentation model to classify and segment the disease spots on wheat leaves. Their method however needs to have disease spots on leaf images accurately segmented. Afridi *et al*. [20] used K-means in plant image processing. However, the k value needs to be specified in advance and the clusters do not have semantic meanings. Zabawa *et al*. [8] proposed a deep-learning-based method for solving the berry detection and counting task. It converts the problem into the pixel-wise classification task of "berry", "edge" and "background", solved by the hourglass encoder-decoder deep network. Lin *et al*. [21] proposed a U-net based convolutional neural network to segment the powdery mildew on cucumber leaf images at the pixel level. However, these existing works need pixel-level annotation, and as it is time-consuming, the collected and labeled dataset is quite small. Despite success in these tasks, image segmentation of lesion areas, which is crucial for quantitative assessment of disease severity, has not been well addressed.

A big challenge to apply deep neural network models in lesion segmentation is that a large number of labeled training images are necessary. Moreover, semantic segmentation typically requires pixel-level or bounding-box image annotation, which needs huge manpower consumption. Due to the scattered distribution of lesion areas, constructing pixel-level labeled datasets for lesion segmentation is even more laborious. Current publicly available plant disease datasets only have disease category labels. Weakly supervised learning, which widely used in various pattern recognition applications including object detection, image segmentation, etc., is suitable for our tasks. It alleviates the need for large-scale detailed annotations and only relies on weaker labels such as image-level category labels, which has profound significance for solving the bottleneck in lesion segmentation.

Supervised semantic image segmentation relies on fully annotated pixel-level training data, which is a major bottleneck for many real-world tasks. To tackle this problem, weakly supervised training methods have been proposed to reduce the annotation effort. There are many ways of weak supervision and the most common

setting is image-level labels, which indicate what kinds of objects are in the image. Previous work [22] demonstrates that the same network can perform both scene recognition and object localization in a single forward pass. Some studies in agriculture [3,7] have also visualized the inner convolutional layers and found that the classification neural networks can efficiently activate the diseased spots on leaf image. In [23], the class activation map (CAM) for CNNs was proposed, in which global average pooling (GAP) and global max pooling (GMP) are used to highlight the most discriminative image regions relevant to the predicted result. This enables CNNs originally trained for classification to learn object location information without using any bounding box or pixel-level annotations. Inspired by the simple yet effective idea of CAM, many practical problems have been solved. Wang *et al*. [24] utilized the CAM method to segment food images for image-based dietary assessment and management. They proposed a new global pooling layer as a cascade combination of a GMP layer and a GAP layer. Bolano *et al*. [25] trained a binary food/non-food classification network to produce a food activation map and then recognized the food type of each candidate. Gondal *et al*. [26] used the CAM idea to segment diabetic retinopathy lesions in retinal fundus images. The CAM method cannot directly get the final segmentation results. Therefore, some post-processing operations are needed, such as threshloding or graph cut. Alternatively, CAM can also be used to produce initial seed regions for a subsequent neural network. For example, Kolesnikov *et al*. [27] expanded the seed regions generated from CAM using a segmentation CNN and then constrained the result to object boundaries using conditional random fields (CRF). These three principles — seed, expand and con-strain (SEC) — have become the classic scheme of weak supervision and a lot of follow-up work (*e.g.*, [28–30]) is carried out on this basis.

In this paper, we propose a novel weakly-supervised segmentation method for crop lesion regions. Firstly, we construct a two-branch network to simultaneously perform classification and generate seed regions (1) in the multi-class classification branch, every single disease type is distinguished, and (2) in the binary classification branch, multiple disease types are classified into a general disease class, and then further processed to generate disease and healthy seed regions for the next segmentation stage. This approach helps ensure disease seeds better cover all lesions, as different disease types often share similar characteristics and differentiating them may lead to certain lesions to be misclassified and missed out in the disease seeds. Since legions have distinctive characteristics of being small and scattered, we utilize an encoder-decoder network with high-resolution output as our segmentation network. Different from previous work (such as SEC [27], DSRG [28]) that use an encoder in the segmentation network, our design to use an encoder-decoder is critical for our system to successfully segment images with small and scattered regions, which is the major challenge in image-based diagnosis of field diseases. We further propose a novel weakly-supervised training strategy, making use of the extracted seed regions. Therefore, the segmentation network can also learn the features from the seed regions, along with simple yet effective self-supervision. Experimental results show that our system achieves state-of-the-art performance for lesion segmentation.

Our system is effective for segmenting images with small and scattered regions, and therefore, can be extended to general multi-class segmentation problems with weak supervision. By additionally introducing a fast conditional-random-field (CRF) supervision inspired by the seed-expand-and-constrain (SEC) method [27], we demonstrate our method on the PASCAL VOC dataset [31]. The results show that our method performs better in images that have similar characteristics of lesion, so has comparable performance with the state-of-the-art DSRG (deep seeded region growing) method [28] in the VOC dataset.

Our main contributions include:

• For lesion seed region detection, we propose an efficient multi-task two-branch network for simultaneous classification and lesion seed region extraction. To ensure robustness, we introduce a multi-label strategy with which subtle types of diseases are grouped into a general disease class in the binary classification branch (in order to generate good disease and healthy seed regions for the latter semantic segmentation network), while in the multi-class classification branch, the subtle classes of different diseases are distinguished.

• We propose a simple yet effective training strategy for weakly-supervised segmentation in the encoder-decoder network structure. The strategy not only significantly reduces the training time, but also achieves better performance on lesion segmentation and other general segmentation problems with similar image structure (*i.e.*, segmenting images with small and scattered regions).

RESULTS AND DISCUSSION

Classification performance

We compare the classification accuracy of our method and Aravind's method [3] in Table 1. We report the classification accuracy of three disease categories and the healthy category. The 3 diseases are black measles,

black rot and leaf blight respectively. Leafs affected by black measles disease often have dark regions with similar distribution as measles. Leafs affected by black rot disease often have brown patches with black fringe. Leafs affected by leaf blight disease often have brown-black leaf spots which look dry. Some example images of the 3 disease types are illustrated in Figs. 2–4. To the best knowledge of the authors, reference [3] is the only method that classifies grape diseases as our system does. Reference [3] is also a state-of-the-art method (published in 2019). Our two-branch network shows powerful classification ability: we get 100% accuracy of healthy vs. disease classification, and 99.63% accuracy for disease type classification, which outperforms Aravind's method (97.62%).

Segmentation performance on PlantVillage

Since no previous work did crop disease segmentation, we choose the classic K-means segmentation as the base-line. Then we compare it with DAM generated from our network combined with CRF [32]. We also adopt two general approaches SEC [27] and DSRG [28] on this special segmentation task for comparison. We compare with SEC and DSRG since they are two representative methods in general weakly-supervised semantic segmentation, are most related methods to our proposed method and achieve relatively good results on lesion region segmentation. We report mean Intersection-over-Union (IoU) results in Table 2. Our method achieves much better segmentation performance than all other methods in all three categories. This demonstrates that our method is effective for the lesion segmentation problem.

Figure 1 shows qualitative results of lesion segmentation comparing several segmentation methods and ours, on one example for each disease. We observe that, for the K-means method, if lesion regions have obvious dark brown characteristics, they can be accurately segmented. However, if there is not enough difference between lesion region sand other part of the image, the segmentation of K-means is very bad. In addition, there is no semantics for clustering results by K-means. SEC, DSRG and DAM+CRF methods all obtain less accurate segmentation than ours, either missing some lesion regions or extracting inaccurate region boundaries. In comparison, our method can effectively handle different diseases and lesion region sizes, and better copes with illumination differences.

The reasons why our method outperforms SEC and

Table 1 Comparison of classification accuracy on PlantVillage (%)

Scheme	Mean	Blk. measles	Blk. rot	Leaf blight	Healthy
[3]	97.62	94.07	95.31	99.53	100
Ours	99.63	99.28	99.58	100	100

Table 2 Comparison of segmentation performance of different methods (measured by IoU (%)) on PlantVillage dataset

Schemes	Mean	Black measles	Black rot	Leaf blight
K-means	24.24	30.69	25.15	16.87
DAM+CRF	24.86	40.99	21.19	12.40
SEC	25.40	38.03	30.24	7.94
DSRG	25.50	34.95	27.03	16.68
Ours	55.49	61.62	55.83	49.02

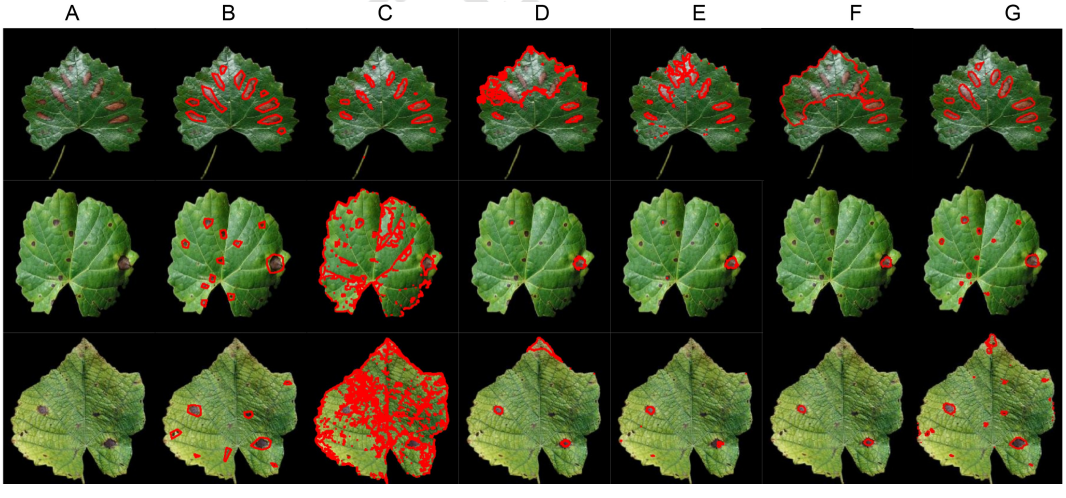

Figure 1. Qualitative results of different methods for lesion segmentation. From left to right: (A) examples of three kinds of grape diseases, (B) ground truth, (C–G) segmentation results of (C) K-means, (D) SEC, (E) DAM+CRF, (F) DSRG, (G) Our method.

DSRG so much are two-fold:

• Firstly, they are not designed for images containing small and scattered regions; in particular, they only use an encoder so that the resolution of output feature map is too low to capture individual regions, whereas we propose to use an encoder-decoder network with a high-resolution output. This design choice is critical to successfully segmenting images with small and scattered regions.

• Secondly, even with the high-resolution output from the encoder-decoder network, DSRG still does not work in practice because the growth of seed regions and the CRF constraint dramatically increase the computational complexity due to the size of the feature map. In our system, we propose pixel-level self-supervision and fast CRF supervision to solve this problem.

More qualitative results of DSRG and our method are presented in Figs. 2– 4. Each figure corresponds to a kind of grape disease. As shown in Figs. 2 and 4, DSRG tends to extract inaccurate region boundaries for grape black measles disease and grape leaf blight disease. As shown in Fig. 3, for grape black rot disease, DSRG easily misses some lesion regions. In comparison, our method generates accurate lesion segmentation on different diseases and under different lighting conditions.

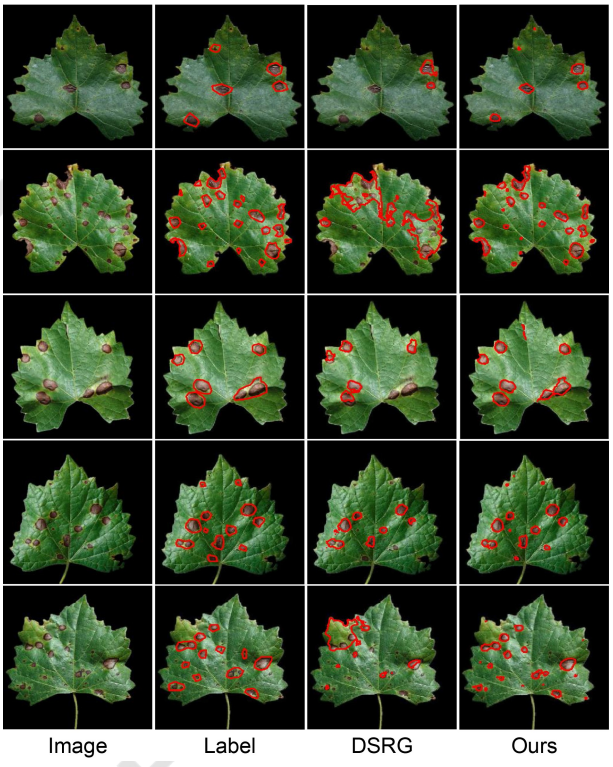

Figure 3. More qualitative results of lesion segmentation on grape black rot disease.

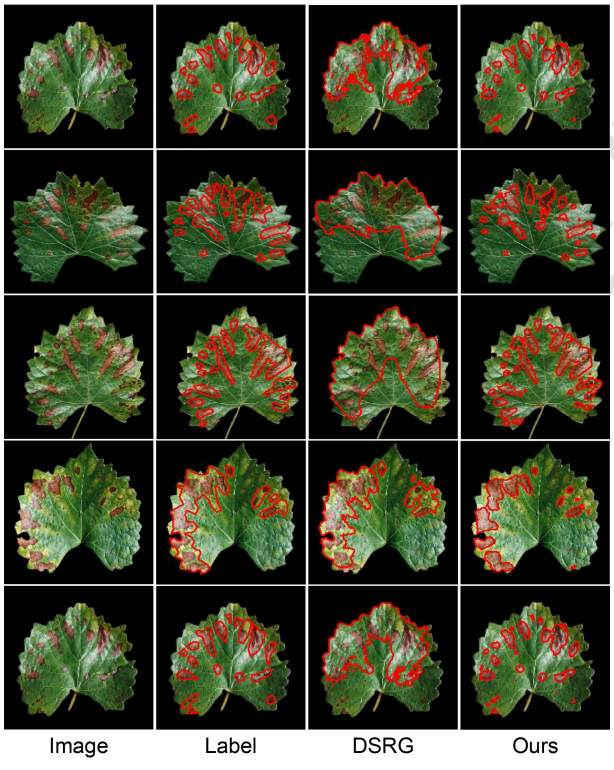

Figure 2. More qualitative results of lesion segmentation on grape black measles disease.

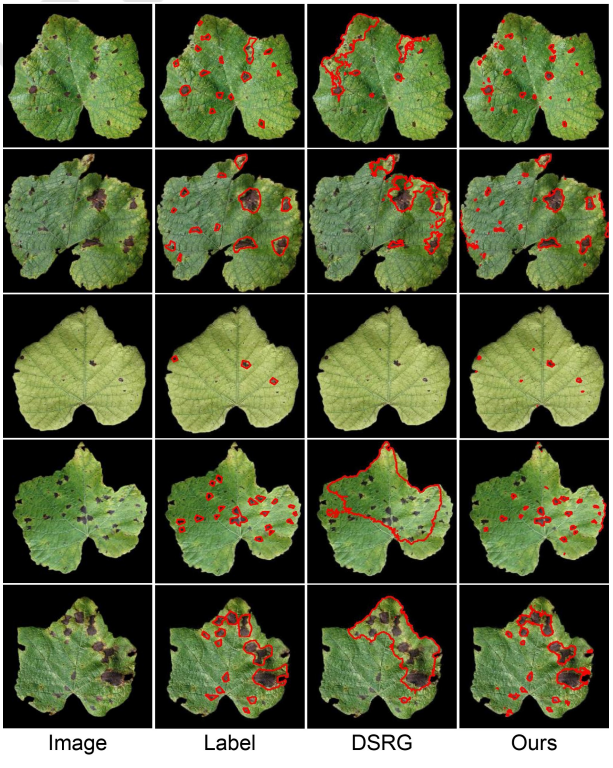

Figure 4. More qualitative results of lesion segmentation on grape leaf blight disease.

Ablation studies

To analyze the effect of each component in our proposed network, we cancel one component at a time compared to the complete proposed network. As shown in Table 3, PSS greatly improves the mean IoU from 26.44% to 55.49%, and the multi-label classification technique has a significant influence in segmentation of class leaf blight, which improves the IoU of leaf blight from 12.33% to 49.02%.

Time efficiency

Another advantage of our proposed new supervision for the encoder-decoder network is that it has better computational efficiency. Compared to the seed loss proposed by DSRG [28] and the CRF loss used by both DSRG [28] and SEC [27], our method can greatly reduce training time. As shown in Table 4, on the network of encoder-decoder structures, using DSRG's seed region growing (SRG) and full-channel CRF will make the training time much longer (approximately 2.4×).

Segmentation performance on VOC

Small and scattered objects are common in natural images. We use the PASCAL VOC dataset as an example to show that in several classes that contain this type of images, our method can achieve good segmentation results.

Per class results of SEC, DSRG and Ours on VOC 2012 validation set are summarized in Table 5. By combining the encoder output trained in a way similar to DSRG with the decoder output trained by our method, our method outperforms DSRG in overall mIoU. In particular, our method works more effectively for images with similar characteristics as lesions, which contributes to the overall performance. Some visual comparisons are presented in Fig. 5, showing that our method outperforms DSRG on small and scattered objects (*e.g.,* part of a chair, upper body of human, small TV).

Comparison with SEC, DSRG and FickleNet [33] on validation and test set are summarized in Table 6. Our method achieves comparable results to the state-of-the-art FickleNet. Although our method does not achieve the

Table 3 Ablation studies on lesion segmentation (in mIoU (%))

Schemes	Mean	Black measles	Black rot	Leaf blight
no PSS	26.44	30.00	26.63	6.69
no LF	43.31	60.21	57.38	12.33
Ours	55.49	61.62	55.83	49.02

No PSS represents not using pixel-level self-supervision. No LF represents not using label fusion but producing seed regions in four-class classification branch.

Table 4 Training time analysis over four loss terms on two networks

Schemes	U-net	DeepLabv3+
PSS	0.44	2.30
PSS+FCS	0.57	4.64
PSS+CRF	0.57	7.86
SRG+FCS	2.55	9.33
SRG+CRF	2.55	11.0

PSS (pixel-level supervision) and FCS (fast CRF supervision) are our proposed supervision. SRG (seed region growing) and CRF are supervision used in DSRG [27]. The results shown are in seconds for each training iteration. The column "DeepLabv3+" shows time analysis of our extension to general segmentation on PASCAL VOC dataset, more settings and results are in Section "Segmentation performance on VOC".

Table 5 Per class results on VOC 2012 validation set, evaluated in terms of mean IoU (%)

Methods	SEC	DSRG (VGG)	DSRG (ResNet)	Ours
Bkg	82.4	87.5	88.0	88.2
Plane	62.9	73.1	78.6	76.5
Bike	26.4	28.4	35.4	36.1
Bird	61.6	75.4	76.2	75.6
Boat	27.6	39.5	42.7	46.4
Bottle	38.1	54.5	62.0	66.8
Bus	66.6	78.2	80.0	81.9
Car	62.7	71.3	68.4	70.6
Cat	75.2	80.6	81.5	81.7
Chair	22.1	25.0	22.6	20.8
Cow	53.5	63.3	77.5	80.5
Table	28.3	25.4	38.6	38.8
Dog	65.8	77.8	73.3	70.9
Horse	57.8	65.4	75.0	74.7
Motor	62.3	65.2	72.6	72.9
Person	52.5	72.8	69.0	69.0
Plant	32.5	41.2	39.5	39.1
Sheep	62.6	74.3	72.8	81.7
Sofa	32.1	34.1	34.9	32.3
Train	45.4	52.1	61.1	60.8
Tv	45.3	53.0	52.2	52.1
mIoU	57.0	59.0	62.0	62.7

The results of DSRG (ResNet) are reproduced by us, whose overall mIoU is slightly higher than the original mIoU 61.4% reported in the paper.

Figure 5. Comparison of DSRG Huang *et al.* (2018) and our method for weakly supervised semantic segmentation on the PASCAL VOC dataset. In these examples, DSRG classified some objects into background or a wrong object. For the examples in rows 1 to 6, the wrong classification is due to that objects appear in incomplete form (unclear reflection of bird, upper body of human, part of a chair, part of a dog, hand, part of a sofa respectively). These objects have the same characteristics as "scattered" lesions. For the examples in the rows 7 to 9, the wrong classification is due to that some objects are small (TV, chair, table respectively). Compared to DSRG, our method can accurately identify and segment these "scattered" and "small" objects.

best performance, as a method designed for images with small and scattered regions, it has already shown its generalization ability to general small and scattered objects besides lesions (as illustrated in Fig. 5).

Table 6 Comparison on VOC 2012 validation and test set, evaluated in terms of mean IoU (%)

Methods	Backbone	Val set	Test set
SEC	VGG16	50.7	51.7
DSRG (VGG)	VGG16	59.0	60.4
DSRG (ResNet)	DeepLab-v2-ResNet101	60.2	63.2
FickleNet	DeepLab-v2-ResNet101	64.9	65.3
Ours	DeepLab-v3-ResNet101&Decoder	62.7	62.6

The result of DSRG (ResNet) on val set is reproduced by us, whose overall mIoU is slightly higher than the original mIoU 61.4% reported in the paper. We show the top-2 methods in val and test set in bold.

CONCLUSIONS

We propose an efficient method to solve the problem of weakly-supervised segmentation of lesions. First we design a two-branch network for multi-task learning: a {*healthy*, *diseased*} two-class branch trained to get the seed regions, and the other branch trained to obtain subtle disease classification. Using the seed regions, we propose a new weakly-supervised segmentation network. The similarity learned by the network allows it to strengthen the features by itself, which forms the pixel-level self-supervision. When applying the method to multi-class segmentation tasks, our proposed fast CRF supervision helps to constrain the prediction of the network in an effective way. Our method not only outperforms state-of-the-art semantic segmentation methods by a large margin for the lesion segmentation task, but also shows its capability to perform well on the PASCAL VOC dataset.

So far our method needs samples of different diseases annotated with category labels for training, which are difficult to obtain. One future work is to extend the current method so that it can be trained using either a small number of samples annotated with category labels or samples without category labels. Another direction of future work is to extend our method to more general application scenarios, for example, dealing with more types of crops, more susceptible locations besides leaf, and more kinds of lesion other than black measles, black rot and leaf blight.

MATERIALS AND METHODS

Our proposed deep model for lesion region segmentation includes two parts. One is a two-branch network used for classifying crop diseases and generating seed regions (Section "Classification and seed region generation"). The other is an overall segmentation

network based on seed regions (Section "Weakly-supervised segmentation network"). The strategies of pixel-level self-supervision and fast CRF supervision are presented in Section "Training with new supervisions".

Classification and seed region generation

Previous works have shown that in addition to being effective for image classification, CNNs also have remarkable ability to localize objects [22,34], where certain image regions can be activated for objects of specific classes. For crop disease classification, some studies also showed this property by visualizing the active regions in the convolutional layer of the classification network [7,35]. These works motivate us to segment lesion regions in crop images based on the classification network.

Previous weakly-supervised segmentation methods (*e.g.*, [28,29]) mainly use the class activation map (CAM) [23] for seed region generation. In more details, these methods train a network to classify the input image into different categories and locate seed regions from obtained CAM. These seed regions are then used by a subsequent network to learn the intrinsic features of different objects. However, our lesion region segmentation task is substantially different from general segmentation tasks: (1) The appearance of different crop diseases are often similar. (2) The fine characteristics of the same type of diseases may have subtle difference. (3) Infection areas on a leaf are often not uniform in number, inconsistent in size (often quite small) and scattered in distribution. (4) Infected and healthy regions may not have clear boundaries. (5) The existence of leaf veins makes the texture information of leaf surfaces complicated.

The above observations pose significant challenges for previous methods based on seed region generation and region expansion. In our work, we propose the following framework to address these issues. In addition to generating lesion seed regions as previous work does, we also generate the healthy seed regions to help with the weakly supervised learning.

Since lesion regions tend to be small and scattered and the healthy region is relatively large, we need to make the disease seed region reach more lesion regions, and make the healthy seed region to avoid all lesion regions. In order to do so, we need a coarse segmentation mask (by thresholding CAM) that covers all lesion regions, so its inversion gives the healthy seed region we want. However, training a classification network as in [23] cannot obtain our desired mask that covers more lesion regions. This is because lesions of different diseases usually have high similarities, and the obtained CAM can only classify them based on the subtle differences of different diseases, which is insufficient to cover the entire area of all lesions. To solve this problem, we make use of a multi-label learning, which assigns two labels to each input image:

$$\begin{aligned} label_1 &\in \{healthy, disease_1, ..., disease_n\} \\ label_2 &\in \{healthy, disease\} \end{aligned} \tag{1}$$

where n is the number of disease categories.

We design a two-branch network to solve this multi-label learning problem, as illustrated in Fig. 6. One branch is a multi-class classification branch trained with $label_1$, to categorize diseases. The other branch is a binary classification branch trained with $label_2$, classifying in two categories (healthy or disease) to avoid missing information. The healthy and disease categories are further used for generating the healthy and disease seed regions. These two branches are trained end-to-end.

In CAM [23], the VGG-16 network [15] was modified by replacing the layers after *conv*5_3 (from *pool*5 to *prob*) with a flat convolution layer, a GAP layer, a fully-connected layer and a softmax layer, to maintain a higher resolution of feature map. Some other works [3,7] visualized activations of the first few convolutional layers and found that certain areas of lesion regions are activated, showing that a few convolutional layers are enough to get activation locations. Meanwhile, as a CNN becomes deeper, the classification performance is improved, meaning that the active region will become more targeted and smaller. So for the purpose of generating an overestimated mask that overlaps all lesion regions, we prefer a shallower CNN to get a better coarse mask. We modify the VGG-16 based network in [23] as follows:

- We use the layers before *conv*4_3 as the basis of feature extraction, followed by two branches.
- For the multi-class classification branch, we use three flat convolution layers with 3 × 3 kernels, three fully connected layers and a softmax layer.
- For the binary classification branch, we use fewer convolution layers, and healthy-disease binary category classification with $label_2$, to obtain a better coarse mask. Unlike detailed classification in [23], this branch includes two flat convolution layers with 3 × 3 kernels (*fc*6_*CAM*, *fc*7_*CAM*), followed by a GAP layer, a fully-connected layer and a softmax layer.

Similar to CAM [23], we extract the disease activation map (DAM) that can be represented as a heat map DAM(x, y) by assigning a heat value (indicating its probability of belonging to the diseased region) to each image pixel. The heat value is computed by the weighted sum of the feature maps from the final convolution layer only from the disease class (the bottom row in Fig. 6):

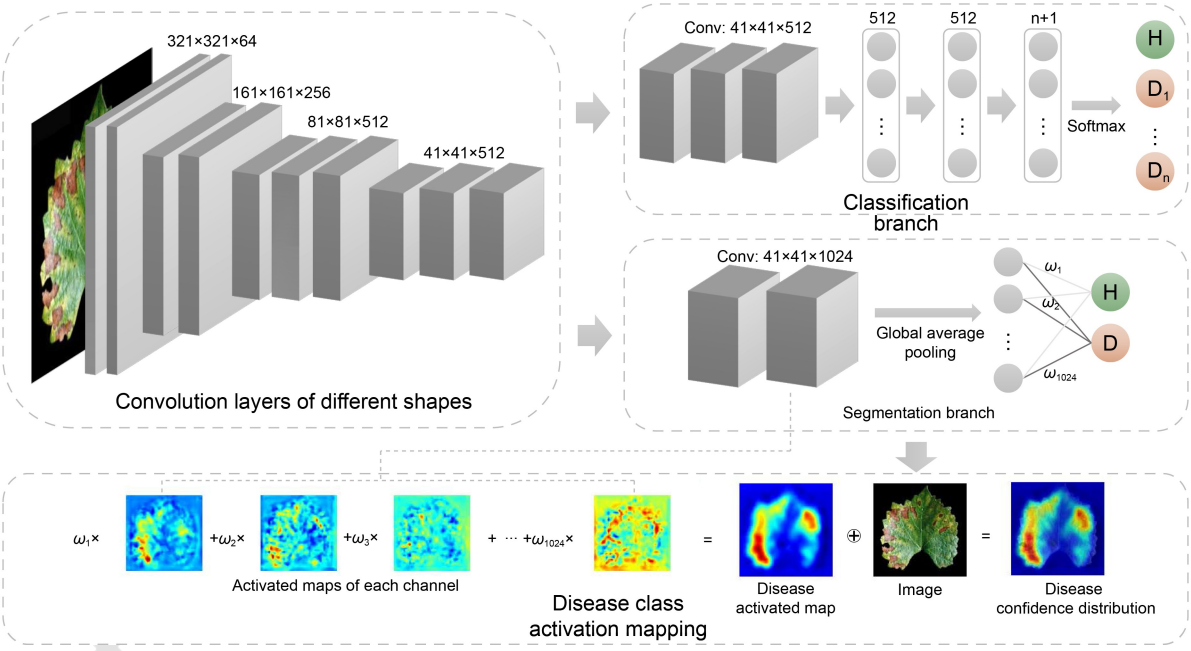

Figure 6. The two-branch network for crop disease classification and seed region generation, which is based on pre-trained VGG-16 network. *H* represents the healthy class, and *D* represents the disease class. The multi-class classification branch provides the prediction for every disease class and similar to CAM [23], the binary classification branch generates the disease and healthy seed regions (see Fig. 7 for some examples).

$$DAM(x,y)=\sum_{k}\omega_k^d\cdot F_k(x,y) \tag{2}$$

where $F_k(x,y)$ represents the k-th feature map of the final convolutional layer (*fc*7_*CAM*) in the binary classification branch, and ω_k^d is the weight learned from the fully-connected layer corresponding to the disease class. We then follow [27,28] to apply a simple thresholding scheme for segmenting the heat map *DAM* into disease and healthy seed regions. The disease seed regions are specified by the values in the heat map that are higher than T_d times the maximum value $d_{\max}$ in the heat map. The healthy seed regions are specified by the values in the heat map that are smaller than T_h times $d_{\max}$. In all our experiments, we fix T_d=0.98 and T_h=0.45.

To handle the dispersed characteristic of lesions, we apply an iterative erasing strategy similar to [36], *i.e.*, erasing is performed M = 5 times. At the m-th erasing, $0\leqslant m\leqslant M$, the m-th disease seed region S_{d_m} is obtained by:

$$S_{d_m}(x,y)=\begin{cases}1, & DAM_m(x,y)>T_d\cdot d_{\max}\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

where DAM_m is the DAM after m-th erasing. The whole lesion seed region S_d is obtained as the union of individual regions:

$$S_d(x,y)=S_{d_0}(x,y)\cup S_{d_1}(x,y)\cup\cdots\cup S_{d_M}(x,y) \tag{4}$$

The whole healthy seed region is consistently specified from DAM_0 by

$$S_h(x,y)=\begin{cases}1, & DAM_0(x,y)>T_h\cdot d_{\max}\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

Some examples of generated whole disease and healthy seed regions are shown in Fig. 7.

Weakly-supervised segmentation network

Similar to previous work [27,28], we use the seed regions to train the weakly-supervised segmentation network. However, in previous work like DSRG [28], seed regions are expanded from seed points by region growing. In our application scenario, for small and scattered objects such as lesions, if a seed is not generated in each lesion, the training may lead to large errors. This is because seed points cannot grow from one lesion to another disconnected lesion. The erasing strategy summarized in Eq. (3) can alleviate this problem to some extent but cannot solve it completely. Meanwhile, the low resolution of the output feature map in these previous methods is unable to segment the lesion region clearly, which is also a bottleneck to be overcome.

To address the above challenges, we propose to use an encoder-decoder network with a high-resolution output $\tilde{F}\in\mathbb{R}^{W\times H\times(C+1)}$, where $\tilde{F}(x,y,c)$ represents the softmax output for each pixel (x,y) belonging to each category c.

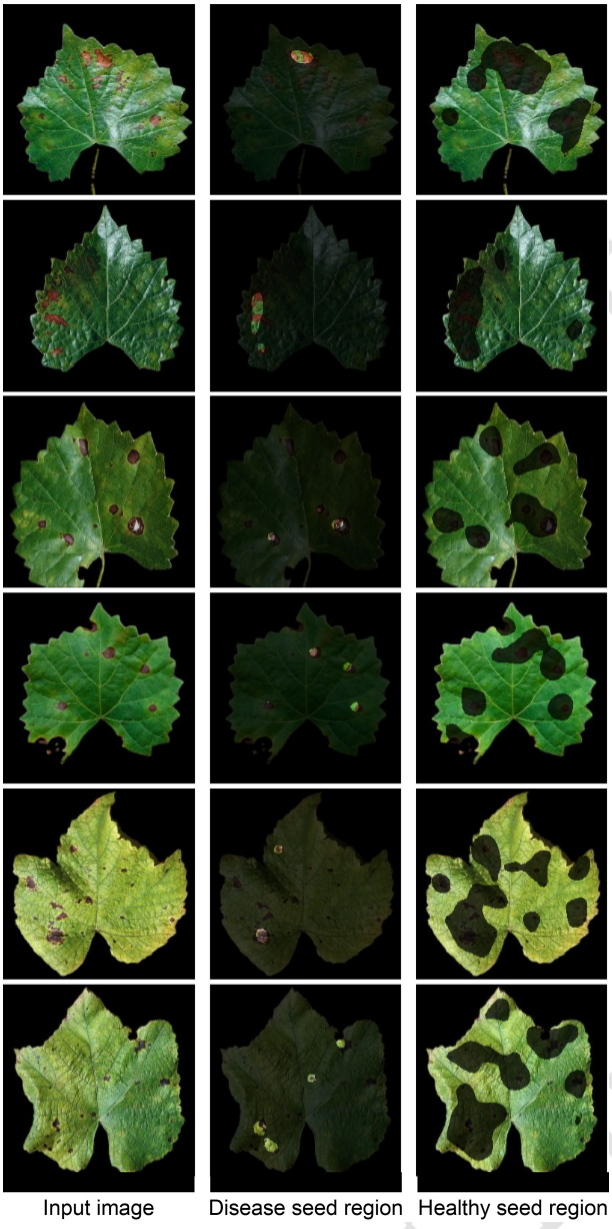

Figure 7. Some examples of disease and healthy seed regions generated by DAM in the two-branch network illustrated in Fig. 6.

Here, W and H are the width and height of the image, C is the number of foreground categories where $c = 1, 2, ..., C$ refers to one of these classes, and an extra class $c = 0$ means the background. For lesion segmentation, $C = 1$ as the foreground corresponds to lesion regions.

Even with the high-resolution output $\tilde{F}$, the previous DSRG method still cannot work since the growth of theseed region [28] and the CRF constraint [27] in DSRG will dramatically increase the computational complexity dueto the size of the feature map, which becomes unacceptable in practice. To solve this problem, we propose two new supervisions, namely pixel-level self-supervision (PSS) and fast CRF supervision (FCS). We then train the network with existing lesion region supervision [28] and the two new supervisions. These two supervisions can not only make the training proceed smoothly, but also make the segmentation network learn the lesion features well. The detailed training scheme is presented in Section "Training with new supervisions". The proposed framework is both applicable to lesion segmentation and can be extended to general segmentation, as illustrated in Fig. 8.

Training with new supervisions

Pixel-level self-supervision (PSS). We propose a simple yet effective scheme for the network to learn from seed regions with self-supervision. We observe that by performing hard thresholding on the network output $\tilde{F}$, we canefficiently achieve the growth of seed regions, which can also extend from one lesion to another, thus addressing the issue with insufficient or incorrect seed region generation. Specifically, the regions for pixel-level self-supervision (PSS) can be obtained automatically as

$$PSS(x,y,c) = \begin{cases} 1, & \widetilde{F}(x,y,c) > \theta_c \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

where θ_c is the threshold. Similar to DSRG [28], we use two different values of θ_c for foreground and background. We then define the pixel-level self-supervision loss l_{PSS} using the method in [28] with the mask obtained using Eq. (6). Network predictions are encouraged to match only the pixels in the supervised region, with the remaining regions ignored.

Fast CRF supervision (FCS). Note that previous methods also impose CRF processing on the network output, which in turn is used as a supervision to constrain the seed growth. However, as the output resolution increases, the CRF computation becomes expensive. To solve this problem, we propose to only use output channels corresponding to object categories present in the input image. Therefore, the input to the CRF module only takes channels corresponding to the categories $c \in I_{\text{tag}}$ where I_{tag} is the set of objects (or background) in the image. Our fast CRF supervision (FCS) mask is then obtained from the output of the CRF module, or0indicating that the corresponding object category does not exist:

$$FCS(x,y,c) = \begin{cases} Output_{\text{crf}}(x,y,c), & \text{if } c \in I_{\text{tag}} \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $Output_{\text{crf}}$ is the output of the CRF module, and FCS is used for the fast CRF supervision. When the

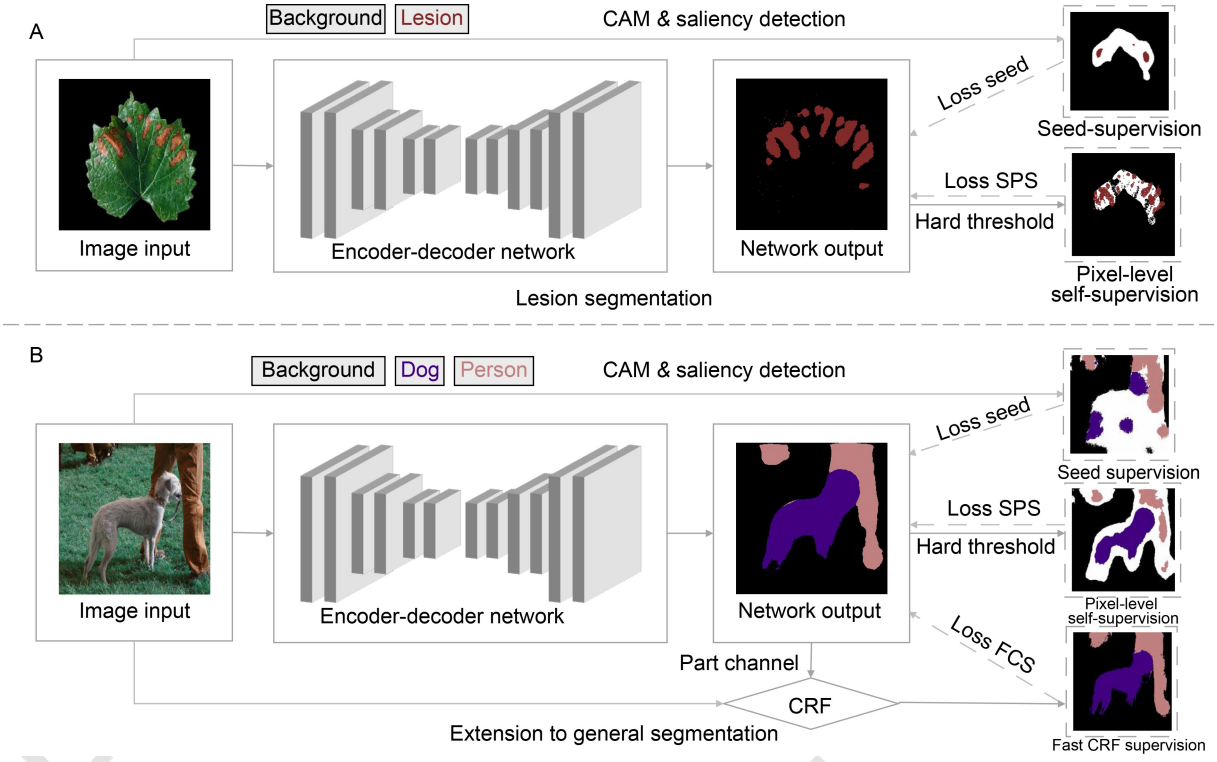

Figure 8. The proposed weakly-supervised segmentation network. We introduce two new supervisions, namely pixel-level self-supervision (PSS) and fast CRF supervision (FCS). For lesion segmentation task, we train the network with seed supervision and PSS. When extending to general segmentation task, we train the network with seed supervision, PSS and FCS.

number of categories is large, the number of channels computed by CRF can be greatly reduced, and then the computational efficiency is improved. In addition, it also corrects the misclassification of the segmentation network. We define the fast CRF loss l_{FCS} using the constraint loss computation method proposed in [27].

Final loss. The final loss function consists of three loss terms: the seeding loss term l_{seed} (adapted from [28] which is the seed region constraint), the PSS loss term l_{PSS}, and the FCS loss term l_{FCS}. The final loss function is formulated as:

$$l = \gamma l_{\text{seed}} + \mu l_{\text{PSS}} + \lambda l_{\text{FCS}} \tag{8}$$

EXPERIMENTAL PROCEDURES

In this section, we validate our proposed method on lesion segmentation task and extend our framework to general segmentation task.

Experimental setup

We implemented the two-branch classification network in Caffe [37] and the weakly-supervised segmentation net-work in TensorFlow [38] on a PC with an NVIDIA GeForce RTX 2080Ti GPU.

Datasets

PlantVillage. The performance of our method is evaluated on a real dataset of grape diseases extracted from the public *PlantVillage* dataset. The images in this dataset have image-level annotation of four different categories (3 diseases + healthy).

The number of images in the training, val and test sets are: (1) healthy: 338 training images, and 85 val images; (2) black measles: 1,107 training images, 277 val images, and 50 test images; (3) black rot: 944 training images, 236 val images, and 50 test images; and (4) leaf blight: 823 training images, 205 val images, and 50 test images; 4,165 images in total.

To quantitatively evaluate lesion region segmentation, a professional research assistant labeled the lesion region of 50 images in the test set for each disease category, 150 in total, using LabelMe. The labeling principle is to label not only those obvious regions with dark brown color, but also those regions with fading and yellowing around those regions with high susceptibility, and to be as careful and accurate as possible. Some examples of labeling are shown in Fig. 1.

Training and testing settings

For the two-branch classification network, we design the network based on modified VGG-16 network in [23]. For the weakly-supervised lesion segmentation network, we use a U-net [39] as the encoder-decoder network for lesion segmentation.

We set θ_c to 0.95 and 0.85 for foreground and background to produce PSS in Eq. (6) in U-net training. In Eq. (8), we set $\gamma = epoch/(epoch+1)$ and $\mu = 1-\gamma$ in U-net, where *epoch* is the current number of times that the training dataset has passed the network (*i.e.*, at the beginning the weight for seeding loss is 0 and the weight for self-supervision loss is 1). In order to match this coefficient setting, we make special treatment for PSS, taking $PSS_{new} = PSS_{old} \cup S_d$ as the supervision. In the lesion weakly supervised segmentation network, the FCS in Eq. (7) is not necessary, so we set $\gamma = 0$.

Extension to general segmentation on VOC

To test the generalization ability of our method, we also apply our method to general scattered object segmentation. We demonstrate this by extending our framework to work on the PASCAL VOC dataset, which achieves comparable performance with the state-of-the-art methods (DSRG [28], FickleNet [33]).

Dataset

PASCAL VOC12. The PASCAL VOC dataset contains 20 object classes and background. For the segmentation task, the training set contains 1,464 images, the validation set contains 1,449 images and the test set contains 1,456 images. Following the common practice in [28, 40,], we use additional training images from [41] to expand the training set to 10,582 images. Although these images have pixel-level annotations, we only use image-level labels in our method for training.

We regard images containing scattered regions of varying sizes as an important type of natural images. So although hour method is developed for lesion region segmentation, we use the general PASCAL VOC dataset as an example to show that in several classes that contain this type of images, our method can achieve good segmentation results.

Training and testing setting

When extending lesion segmentation to general VOC segmentation, we use ResNet101 [16] based DeepLabv3+ [40] (instead of U-net [39]) as the encoder-decoder network for general semantic segmentation (VOC). The above different network choices are due to the different numbers of categories in the two datasets. For lesion segmentation, we prefer to build a network that is fast and efficient in both training and prediction. For the VOC dataset, it contains dozens of object types, so a deeper network with large learning capabilities is needed. The seeds are produced by [28], which are further post-processed by CRF.

The training strategy is also different: on the VOC dataset, we set all the three loss coefficients in Eq. (8) to 1.We first train the encoder network in a similar way as DSRG, since the feature map output from the encoder is of low resolution. Then the encoder parameters are frozen, and the decoder is trained using our proposed method. We use both the output of encoder and the output of decoder to predict pixel-level pseudo-tag (as no pixel-level ground truth is used), and select one of the outputs based on the classification accuracy. Following the previous work [28,30], we use a complete DeepLabv3+ network to retrain on pseudo-tag in a fully supervised way to improve results.

ACKNOWLEDGEMENTS

This work was partially supported by the National Natural Science Foundation of China (Nos. 61725204 and 62002258) and a Grant from Science and Technology Department of Jiangsu Province, China.

COMPLIANCE WITH ETHICS GUIDELINES

The authors Ran Yi, Rui Zeng, Yang Weng, Minjing Yu, Yu-Kun Lai and Yong-Jin Liu declare that they have no conflict of interest or financial conflicts to disclose.

REFERENCES

1. Strange, R. N. and Scott, P. R. (2005) Plant disease: a threat to global food security. Annu. Rev. Phytopathol., 43, 83−116
2. Geiger, F., Bengtsson, J., Berendse, F., Weisser, W. W., Emmerson, M., Morales, M. B., Ceryngier, P., Liira, J., Tscharntke, T., Winqvist, C., *et al*. (2010) Persistent negative

effects of pesticides on biodiversity and biological control potential on european farmland. Basic Appl. Ecol., 11, 97−105
3. Aravind, K. R., Raja, P., Aniirudh, R., Mukesh, K. V., Ashiwin, R. and Vikas, G. (2018) Grape crop disease classification using transfer learning approach. In: Proc. ISMAC-CVB, pp. 1623–1633
4. DeChant, C., Wiesner-Hanks, T., Chen, S., Stewart, E. L., Yosinski, J., Gore, M. A., Nelson, R. J. and Lipson, H. (2017) Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning. Phytopathology, 107, 1426−1432
5. Pound, M. P., Atkinson, J. A., Wells, D. M., Pridmore, T. P. and French, A. P. (2017) Deep learning for multi-task plant phenotyping. In: Proc. ICCV Workshops, pp. 2055–2063
6. Abdu, A. M., Mokji, M. M. and Sheikh, U. U. (2019) Deep learning for plant disease identification from disease region images. In: Proc. ICIRA, pp. 65–75
7. Mohanty, S. P., Hughes, D. P. and Salathé, M. (2016) Using deep learning for image-based plant disease detection. Front Plant Sci, 7, 1419
8. Zabawa, L., Kicherer, A., Klingbeil, L., Milioto, A., Topfer, R., Kuhlmann, H. and Roscher, R. (2019) Detection of single grapevine berries in images using fully convolutional neural networks. In: Proc. CVPR Workshops
9. Zhang, S., You, Z. and Wu, X. (2019) Plant disease leaf image segmentation based on superpixel clustering and EM algorithm. Neural Comput. Appl., 31, 1225−1232
10. Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks, In: Proc. NeurIPS, pp. 1097–1105
11. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going deeper with convolutions. In: Proc. CVPR, pp. 1–9
12. Krause, J., Baek, K. and Lim, L. (2019) A guided multi-scale categorization of plant species in natural images. In: Proc. CVPR Workshops
13. Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. C. and Soares, J. V. (2012) Leafsnap: A computer vision system for automatic plant species identification. In: Proc. ECCV, pp. 502–516
14. Fuentes, A., Yoon, S., Kim, S. C. and Park, D. S. (2017) A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. Sensors (Basel), 17, 2022
15. Simonyan, K. and Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition, In: Proc. ICLR
16. He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In: Proc. CVPR, pp. 770–778
17. Chen, Y., Baireddy, S., Cai, E., Yang, C. and Delp, E. J. (2019) Leaf segmentation by functional modeling. In: Proc. CVPR Workshops
18. Phadikar, S., Sil, J. and Das, A. K. (2013) Rice diseases classification using feature selection and rule generation techniques. Comput. Electron. Agric., 90, 76−85
19. Johannes, A., Picon, A., Alvarez-Gila, A., Echazarra, J., Rodriguez-Vaamonde, S., Navajas, A. D. and Ortiz-Barredo, A. (2017) Automatic plant disease diagnosis using mobile capture devices, applied on a wheat use case. Comput. Electron. Agric., 138, 200−209
20. Afridi, M. J., Liu, X. and McGrath, J. M. (2014) An automated system for plant-level disease rating in real fields. In: Proc. ICPR, pp. 148–153
21. Lin, K., Gong, L., Huang, Y., Liu, C. and Pan, J. (2019) Deep learning-based segmentation and quantification of cucumber powdery mildew using convolutional neural network. Front Plant Sci., 10, 155
22. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. and Torralba, A. (2015) Object detectors emerge in deep scene cnns. In: Proc. ICLR
23. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. and Torralba, A. 2016. Learning deep features for discriminative localization. In: Proc. CVPR, pp. 2921–2929
24. Yu, W., Zhu, F., Boushey, C. J. and Delp, E. J. (2017) Weakly supervised food image segmentation using class activation maps. In: Proc. ICIP, pp. 1277–1281
25. Bolaños, M. and Radeva, P. (2016) Simultaneous food localization and recognition. In: Proc. ICPR, pp. 3140–3145
26. Gondal, W. M., Kohler, J. M., Grzeszick, R., Fink, G. A. and Hirsch, M. (2017) Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In: Proc. ICIP, pp. 2069–2073
27. Kolesnikov, A. and Lampert, C. H. (2016) Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Proc. ECCV, 695–711
28. Huang, Z., Wang, X., Wang, J., Liu, W. and Wang, J. (2018) Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proc. CVPR, pp. 7014–7023
29. Wang, X., You, S., Li, X. and Ma, H. (2018) Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proc. CVPR, pp. 1354–1362
30. Ahn, J. and Kwak, S. (2018) Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proc. CVPR, pp. 4981–4990
31. Mottaghi, R., Chen, X., Liu, X., Cho, N. G., Lee, S. W., Fidler, S., Urtasun, R. and Yuille, A. (2014) The role of context for object detection and semantic segmentation in the wild. In: Proc. CVPR, pp. 891–898
32. krähenbühl, P. and Koltun, V. (2011) Efficient inference in fully connected crfs with gaussian edge potentials. In: Proc. NeurIPS, pp. 109–117
33. Lee, J., Kim, E., Lee, S., Lee, J. and Yoon, S. (2019) Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In: Proc. CVPR, pp. 5267–5276
34. Oquab, M., Bottou, L., Laptev, I. and Sivic, J. (2015) Is object localization for free? – Weakly-supervised learning with convolutional neural networks. In: Proc. CVPR, pp. 685–694
35. Liu, B., Zhang, Y., He, D. and Li, Y. (2018) Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry (Basel), 10, 11

36. Chaudhry, A., Dokania, P. K. and Torr, P. H. (2017) Discovering class-specific pixels for weakly-supervised semantic segmentation. arXiv, 1707.05821
37. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R. B., Guadarrama, S. and Darrell, T. (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proc. MM, pp. 675–678
38. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., *et al*. (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv, 1603.04467
39. Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI, pp. 234–241
40. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proc. ECCV, pp. 801–818
41. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S. and Malik, J. (2011) Semantic contours from inverse detectors. In: Proc. ICCV, pp. 991–998