applied sciences

MDPI

Article

Leaf Spot Attention Networks Based on Spot Feature Encoding for Leaf Disease Identification and Detection

Chang-Hwan Son

Department of Software Convergence Engineering, Kunsan National University, Gunsan 54150, Korea; cson@kunsan.ac.kr

Abstract: This study proposes a new attention-enhanced YOLO model that incorporates a leaf spot attention mechanism based on regions-of-interest (ROI) feature extraction into the YOLO framework for leaf disease detection. Inspired by a previous study, which revealed that leaf spot attention based on the ROI-aware feature extraction can improve leaf disease recognition accuracy significantly and outperform state-of-the-art deep learning models, this study extends the leaf spot attention model to leaf disease detection. The primary idea is that spot areas indicating leaf diseases appear only in leaves, whereas the background area does not contain useful information regarding leaf diseases. To increase the discriminative power of the feature extractor that is required in the object detection framework, it is essential to extract informative and discriminative features from the spot and leaf areas. To realize this, a new ROI-aware feature extractor, that is, a spot feature extractor was designed. To divide the leaf image into spot, leaf, and background areas, the leaf segmentation module was first pretrained, and then spot feature encoding was applied to encode spot information. Next, the ROI-aware feature extractor was connected to an ROI-aware feature fusion layer to model the leaf spot attention mechanism, and to be joined with the YOLO detection subnetwork. The experimental results confirm that the proposed ROI-aware feature extractor can improve leaf disease detection by boosting the discriminative power of the spot features. In addition, the proposed attention-enhanced YOLO model outperforms conventional state-of-the-art object detection models.

Keywords: smart farming; leaf disease identification; leaf disease detection; feature extractor

Citation: Son, C.-H. Leaf Spot Attention Networks Based on Spot Feature Encoding for Leaf Disease Identification and Detection. *Appl. Sci.* **2021**, *11*, 7960. https://doi.org/10.3390/app11177960

Academic Editor: Joonki Paik

Received: 6 August 2021
Accepted: 27 August 2021
Published: 28 August 2021

1. Introduction

Smart farming refers to the management of farms using information and communication technologies to increase the quantity and quality of plants and crops. By placing smart agriculture sensors in greenhouses or in the field, various sensing data such as lighting, temperature, soil nutrient levels, leaf color, and humidity can be collected. Given the vast amount of sensing data, crop growth can be evaluated using data analysis tools to enable farmers to make data-driven decisions. In other words, farmers can determine optimal amounts of water, fertilizers, and pesticides to minimize resources and raise better and healthier crops.

Particularly, crop disease diagnosis in a timely manner is important to prevent diseases from spreading at an immature state and prevent economic damages to farmers. A large team of experts and farmers can identify crop diseases based on the symptoms on the leaves; however, this manual observation is time consuming and costly. In addition, it is inefficient to continuously monitor all the crops on a large field area. Therefore, the automatic detection of crop diseases is necessary.

With the rapid advance in computer vision enabled by deep learning, image-based crop disease detection has garnered particular attention. Among deep learning models, the deep convolutional neural network (DCNN) [1] has demonstrated powerful performance for image classification and detection. Therefore, image-based approaches have been actively studied using digital cameras built on autonomous agricultural vehicles for crop

disease identification and detection. This study only deals with the identification and detection of apple leaf diseases.

1.1. Related Works

1.1.1. Leaf Disease Identification

Conventionally, image-based leaf disease identification involves two steps: image feature extraction and classifier learning. The first step is image feature extraction, which refers to the process of describing the local image appearance for leaf disease detection and the generation of image-level feature vectors. To characterize local leaf spot appearances, popular feature extractions such as SIFT [2], LBP [3], sparse codes [4], and others [5], including color histograms and entropy, can be applied. Subsequently, these features are pooled and aggregated through Bag-of-words (BOW) [6] and Fisher vector encoding (FVE) [7] to obtain image-level feature vectors. The second step is classifier learning to find a hyperplane that can separate image-level feature vectors into classes. Given these feature vectors, a support vector machine (SVM) [8,9], which is a data analysis tool, is trained to classify leaf diseases. Certainly, other tools, such as decision trees [10], genetic algorithms [11], and dictionary learning [12], can be utilized for leaf disease identification.

Recently, the DCNN has replaced a series of steps that consisted of handcrafted feature designs, pooling, and classifier learning, because the DCNN can automatically learn generic representations in a hierarchical manner for discriminative feature extraction. With the emergence of the DCNN, a profound knowledge of feature design, feature pooling, and classifier learning is not necessary, thereby rendering it easier for non-experts to handle the leaf disease identification problems. If a new training dataset is provided, good performance can be obtained through transfer learning (TL), which uses pretrained models such as VGG [13] and ResNet [14], and subsequently updates the model's parameters. A large number of studies [15–21] have been performed based on TL during the past few years for leaf disease identification.

More recently, attention networks [22], feature pyramid networks [23], and vision transformer networks [24] have been actively studied. Attention networks [22] model spatial and channel weighting maps, to emphasize the features in a particular area or channel. Feature pyramid networks [23] utilize multiple feature maps with different scales in the backbone, which refers to general-purpose feature extractors such as VGG and ResNet, to be more robust to the object's scale problem. Vision transformer networks [24] replace the DCNN backbone as a convolution-free model, and employ a pure transformer and pyramid transformer as a unified backbone for various vision tasks. A sequence of patches is adopted as the input, which is different from conventional DCNNs. It has been reported that vision transformer networks can be applied to many downstream tasks, while outperforming traditional backbones.

1.1.2. Leaf Diseases Detection

Conventional object detection approaches can be adopted for leaf disease detection [25]. Unlike leaf disease identification, leaf disease detection requires both the category and location of leaf diseases. Conventionally, two approaches have been adopted for object detection. One is a sliding-window detector [26], and the other is a region proposal [27]. The sliding-window detector moves a window along a raster scanning direction, and determines whether the window contains leaf diseases. During the training phase, handcrafted features such as histogram of gradients (HOG) [26], LBP [3], or SIFT [2] are extracted from positive and negative samples. Subsequently, SVM [9,10] is trained for leaf disease classification. In the test phase, the same feature extractor is applied repeatedly for each sliding window, and the learned SVM determines whether leaf diseases are contained in the window.

The region proposal is a technique for generating candidates (i.e., bounding boxes) where leaf diseases may exist. It starts by generating superpixels [28], and merging them with similar colors, sizes, and textures. Here, a superpixel is defined as a group of pixels that

share common characteristics. Through bottom-up grouping, approximately a thousand bounding boxes surrounding the superpixels before and after merging are generated. Similarly to the sliding-window detector, handcrafted features such as SIFT and its variants are generated for each bounding box, and the SVM classifier is trained. Selective Search [27] and EdgeBoxes [29] are popular methods for region proposals.

Recent object detection technologies use pretrained models such as VGG [13] and ResNet [14] as backbones for feature extractors. For each bounding box generated by a region proposal, the pretrained models extract deep features through a layer-by-layer transformation. The region-based CNN (RCNN) [30], which is the earliest model, adopts this approach. Two stages are required: region proposal and classifier learning. In RCNN, training is a multi-stage pipeline, and object detection is slow. To overcome these drawbacks, a Fast RCNN [31] has been proposed. In this model, a multi-task loss is utilized to jointly minimize class and bounding-box losses, and an ROI pooling layer is utilized to transform each candidate into a fixed-sized feature map. This enables training to be in a single stage. However, the region proposal is not separated from the Fast RCNN. To conduct the region proposal in the feature domain, a region proposal network (RPN) [32] has been proposed.

In addition to the RCNN series, there is another single-shot detector that treats object detection as a single regression problem. YOLO [33] and SSD [34] are popular models. Unlike the RCNN series, YOLO directly predicts bounding boxes from a full image. This creates a high-speed and high-accuracy detector. Similar to YOLO, SSD is a single-shot detector; however, it utilizes multiple features with different scales to handle various objects in size. In addition, the SSD defines a set of default boxes with different aspect ratios, which correspond to anchors in a Faster RCNN [32].

Recently, Mask RCNN [35] and Cascaded RCNN [36] have been introduced. Mask RCNN extends the Faster RCNN by adding a new branch for instance segmentation. Therefore, the Mask RCNN can simultaneously achieve localization and instance segmentation. The Cascaded RCNN deals with a noisy detection problem, due to an intersection over union (IOU) threshold. To address the limitation, the Cascaded RCNN trains a sequence of detectors with increasing IOU thresholds. More details are provided in the literature [36].

1.2. Motivation

To the best of our knowledge, there are few deep-learning models specialized for leaf disease identification and detection. Existing models such as VGG, ResNet, YOLO, and Faster R-CNN are directly utilized, or with minor modifications. In other words, the TL approach dominates leaf disease identification and detection [15–21,37,38]. Unlike popular datasets such as ImageNet [1] and VOC2007 [26], which have been adopted for conventional computer vision problems such as classification and object detection, leaf disease datasets have characteristics in that there is a spatial region in which leaf diseases exist.

Figure 1 illustrates an example of an apple leaf with diseases. Leaf diseases solely exist in the leaves, whereas the background has no information about them. Therefore, it is essential to find the spot area (i.e., the diseased area), and extract spot features to identify leaf diseases. These spot features are informative and discriminative and play a crucial role in leaf disease identification and detection. Based on this observation, this study reveals how to extract informative and discriminative spot features from input leaf images. A novel leaf spot attention network (LSA-Net), and attention-enhanced YOLO (AE-YOLO) network equipped with an ROI-aware feature extractor, are proposed for leaf disease identification and detection, respectively. The proposed networks incorporate a leaf spot attention mechanism to find spot areas and encode spot information.

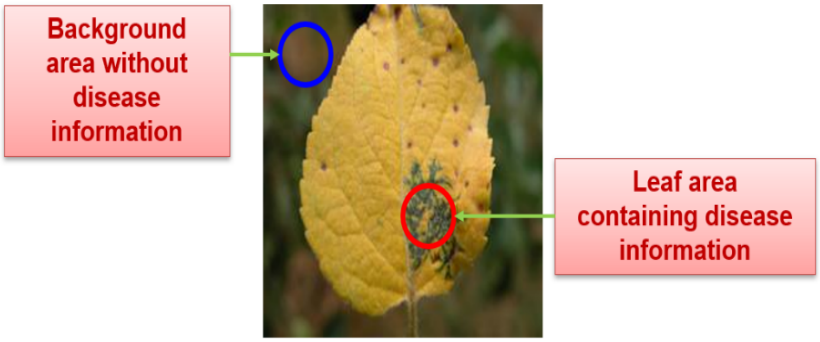

Figure 1. Leaf image characteristics.

1.3. Proposed Approach

Figure 2 depicts the difference between traditional approaches [5,9,39] and the proposed networks for leaf disease identification and detection. As shown in Figure 1, the traditional approach includes three steps, i.e., clustering, feature extraction, and classifier learning. First, the clustering step divides the input image into background and spot areas, and two types of features are extracted from the divided areas via a feature extraction step. Next, the classifier is trained using machine learning tools such as SVM and decision trees. However, this traditional approach has a few drawbacks. The first drawback is that the colors and textures used for clustering and feature extraction are handcrafted features, which are not better than deep features in terms of the discriminative power because deep learning can provide abundant features via a layer-by-layer nonlinear transformation. The second drawback is that the employed classifiers, such as SVM and decision trees, are not better than deep learning in terms of classification performance.

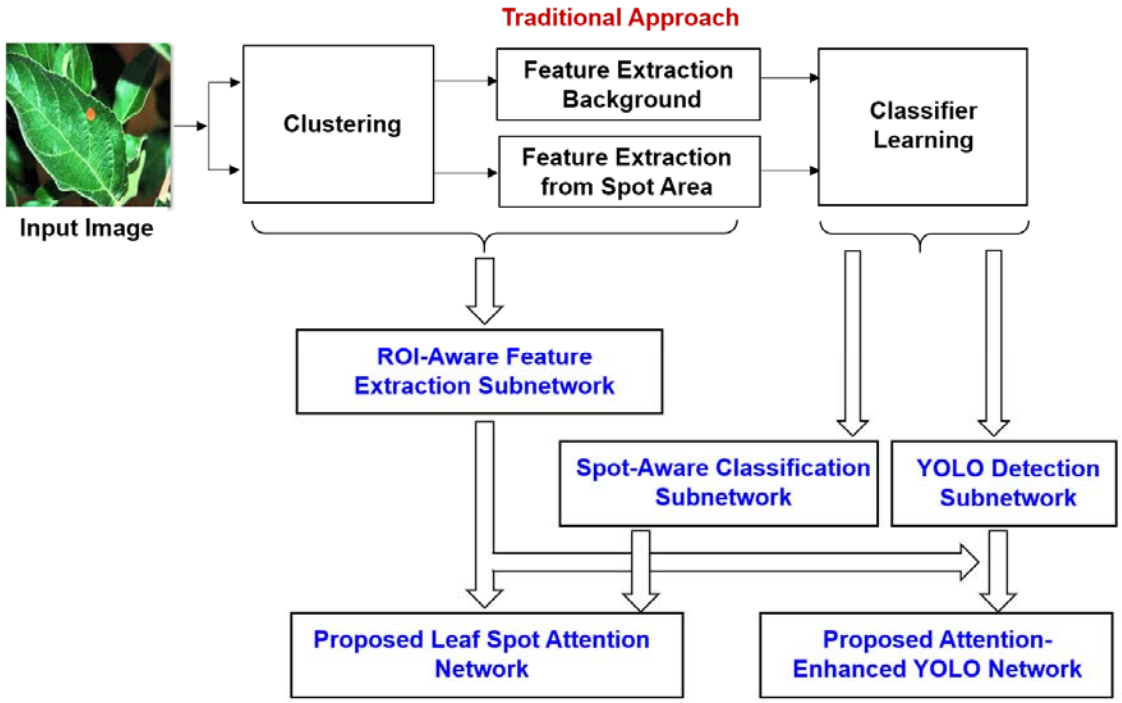

Figure 2. Traditional approach [39] versus proposed networks for leaf disease identification and detection.

To overcome the shortcomings of traditional approaches, the proposed method incorporates the three steps into a single deep learning architecture. Inspired by conventional approaches, the clustering and feature extraction are replaced by the ROI-aware feature

extraction subnetwork (ROI-aware FES), and the classifier learning is replaced by the SAC-SubNet or YOLO detection subnetwork, as illustrated in Figure 2. For leaf disease identification, ROI-aware FES and SAC-SubNet are adopted. The entire network that comprises the ROI-aware FES and SAC-SubNet is referred to as LSA-Net. For leaf disease detection, the ROI-aware FES and YOLO detection subnetwork are adopted, and this detection network is referred to as the AE-YOLO network.

The primary idea of the proposed networks is based on the observation that symptoms of leaf diseases can solely be detected in the leaf area. To realize this, the ROI-aware FES, which is a new spot feature extractor, is designed, and the ROI-aware feature fusion is proposed to model the leaf spot attention mechanism. The ROI-aware FES predicts a leaf segmentation map, and subsequently encodes spot information. The ROI-aware feature fusion combines two features provided by the feature extractors, that is, the backbone and ROI-aware FES. This feature fusion models the leaf spot attention mechanism, and increase the discriminative power. The ROI-aware FES and feature fusion can teach the SAC-SubNet and YOLO detection subnetwork regarding which areas and features have a decisive role in classifying and localizing leaf diseases. Therefore, the proposed ROI-aware FES and feature fusion can be viewed as attention mechanisms [22,24]. Owing to the leaf spot attention enabled by the ROI-aware FES and feature fusion, the proposed networks can improve leaf disease identification and detection.

Although deep learning approaches [15–21,37,38] have been recently introduced, existing models, such as VGG and ResNet, have been directly utilized for leaf disease identification and detection. That is, TL dominates. However, this approach has limitations in improving the discriminative power, because these models exclude the leaf spot attention mechanism to extract informative features from leaf images. This is the major difference between the proposed network and conventional transfer learning-based deep learning.

1.4. Contributions

- This paper is an updated version of the previous work published in the IEEE CVPR workshop [40]. The previous work revealed that the proposed ROI-aware FES is very effective in improving leaf disease identification performance. In this line of research, it is interesting that the proposed ROI-aware FES can also be applied to leaf disease detection. Therefore, in this paper, we illustrate how to incorporate the proposed ROI-aware FES, that is, a new spot feature extractor, into the conventional YOLO framework. The advanced YOLO model that incorporates the ROI-aware FES and feature fusion is referred to as AE-YOLO in this study. In addition, it is revealed that the proposed AE-YOLO can improve leaf disease detection and surpass state-of-the-art object detection models. The proposed AE-YOLO is also expected to be applicable not only for the detection of apple leaf diseases, but also for the detection of pests and diseases in other crops.
- To the best of our knowledge, this study is the first attempt to introduce a novel deep learning architecture that considers the leaf spot attention mechanism and is applicable for both leaf disease identification and detection. Until now, existing models such as VGG, ResNet, and YOLO have been adopted for leaf disease identification and detection [15–21,37,38]. However, in the proposed architecture, a new ROI-aware FES and feature fusion are introduced to find spot areas and encode spot information. The major contribution is to show a novel deep learning architecture that can incorporate the leaf spot attention mechanism based on the ROI-aware FES into existing classification and detection models.
- Previous studies have targeted leaf images with a single background color and a single leaf [17,21]. These images are simple, and good results can be obtained by applying only TL. However, in this study, more complicated images are tested. In other words, the background has few branches and leaves. This is a much more challenging problem. The leaf disease dataset adopted in this study will be open to the public for research purposes.

- This study reveals how to incorporate the conventional approach [5,9,39] (including three steps, i.e., clustering, feature extraction, and classifier learning) into a single deep learning architecture. In the proposed networks, the clustering and feature extraction steps are replaced by the proposed ROI-aware FES, and the classifier learning is replaced by the SAC-SubNet or YOLO detection subnetwork. The major difference between the proposed method and the conventional approach is that the proposed method performs three steps simultaneously. In addition, unlike conventional clustering [39], which might fail to extract spot colors from leaf images because of the similarity in their background and spot colors, the proposed method incorporates the clustering algorithm into the deep learning architecture; thus, more accurate feature clustering can be obtained.

2. Proposed Networks

2.1. Proposed LSA-Net for Leaf Disease Identification

Symptoms can be detected only in the leaf area, whereas the background region contains no information regarding leaf diseases. Therefore, the additional use of the ROI features that contain the leaf, background, and spot information can teach the SAC-SubNet regarding which features are more important and which features should have a decisive role in classifying leaf diseases. Hence, an additional subnetwork to extract the spot features from input leaf images is designed and subsequently combined with the SAC-SubNet.

Figure 3 illustrates the proposed LSA-Net for identifying apple leaf diseases. The proposed architecture consists of two subnetworks: ROI-aware FES and SAC-SubNet. First, the ROI-aware FES is a feature extraction network comprising two modules. One is the leaf segmentation module, and the other is the spot feature encoding module. The leaf segmentation module divides an input leaf image into background, leaf, and spot areas.

$$\boldsymbol{J} = \boldsymbol{f}_{\boldsymbol{\theta}_{seg}}(\boldsymbol{I}) \tag{1}$$

where $f_{\boldsymbol{\theta}_{seg}}$ is the learnable function with the parameter $\boldsymbol{\theta}_{seg}$ for the leaf segmentation module. $\boldsymbol{I}$ and $\boldsymbol{J}$ denote the input leaf image and segmented feature map, respectively. The architecture of the $f_{\boldsymbol{\theta}_{seg}}$ is a fully convolutional network (FCN), developed for semantic segmentation [41]. In the leaf segmentation module, the cropping layer applies two-dimensional cropping to the input feature maps. Two input feature maps are required. One is to be cropped, and the other is the reference to determine the size of the cropped feature map. The transposed convolution layer applies the transposed convolution to the input feature maps for up-sampling, and $\oplus$ indicates the addition layer that adds the input feature maps by element.

The spot feature encoding module maps the segmented feature map $\boldsymbol{J}$ into a spot feature vector $\boldsymbol{v}$.

$$\boldsymbol{v} = f_{\boldsymbol{\theta}_{enc}}(\boldsymbol{J}) \tag{2}$$

where $f_{\boldsymbol{\theta}_{enc}}$ is the learnable function for the spot feature encoding module, and is composed of pooling and fully connected layers. The segmented feature map, $\boldsymbol{J}$, is high-dimensional and contains spatial information. However, the spot feature vector, $\boldsymbol{v}$, is low-dimensional and has no spatial information. This means that the spot information contained in $\boldsymbol{J}$ is encoded into the low-dimensional vector $\boldsymbol{v}$.

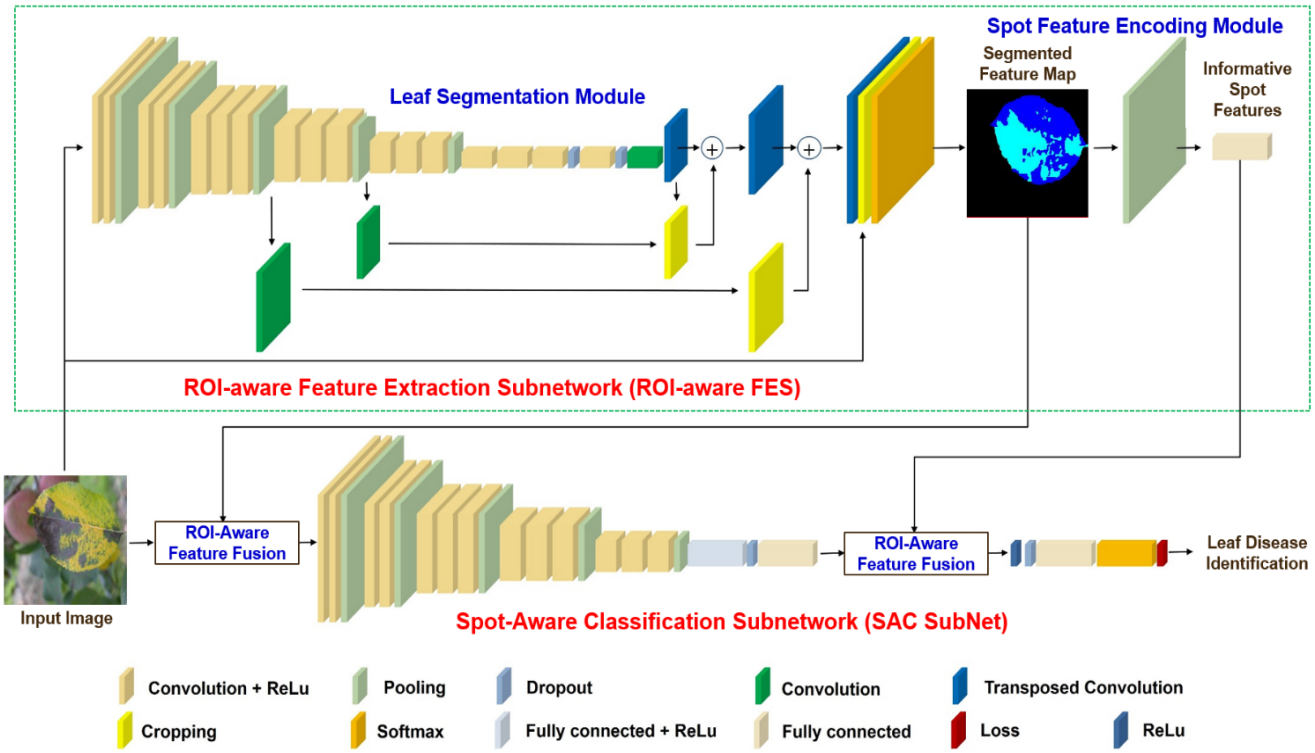

Figure 3. Proposed LSA-Net for leaf disease identification.

Second, the SAC-SubNet is a classification network and is an advanced version of the existing VGG model, equipped with ROI-aware feature fusions to reflect the leaf spot attention mechanism. In SAC-SubNet, there are two ROI-aware feature fusions. One is early fusion, and the other is later fusion. As mentioned in the previous paragraph, the two features $\boldsymbol{J}$ and $\boldsymbol{v}$ have different spatial information and dimensions. Therefore, the features $\boldsymbol{J}$ and $\boldsymbol{v}$ are utilized for early and later fusions, respectively.

$$\boldsymbol{F}^{early} = f_c(\boldsymbol{J}, \boldsymbol{I}) \tag{3}$$

$$\boldsymbol{F}^{later} = f_c\left(\boldsymbol{v}, f^{(i)}_{\boldsymbol{\theta}_{vgg}}\left(\boldsymbol{I}, \boldsymbol{F}^{early}\right)\right) \tag{4}$$

where f_c is the function for the early and later fusions. In the LSA-Net, simple concatenation is adopted for the function, which indicates that f_c has no parameters to be updated. Even though another attention model, such as scaled dot-product attention [42], can be considered, our experiment revealed that concatenation provides better performance than scaled dot-product attention. Early fusion is indicated on the left side of SAC-SubNet. Because the function f_c is defined as the concatenation, the segmented feature map $\boldsymbol{J}$ is stacked on top of the input leaf image $\boldsymbol{I}$. In Equation (3), $\boldsymbol{F}^{early} = [\boldsymbol{I}; \boldsymbol{J}]$ is the early fusion result. Here, a semicolon (;) is utilized to indicate a new row in the matrix. The later fusion is indicated at the center right of the SAC-SubNet. The later fusion concatenates two features $\boldsymbol{v}$ and $f^{(i)}_{\boldsymbol{\theta}_{vgg}}$, as expressed in Equation (4), where $f^{(i)}_{\boldsymbol{\theta}_{vgg}}$ denotes the output feature maps of the ith layer of the VGG model. In this study, the VGG16 model was adopted for SAC-SubNet. In Equation (4), it is noted that $f^{(i)}_{\boldsymbol{\theta}_{vgg}}$ takes two inputs, $\boldsymbol{I}$ and $\boldsymbol{F}^{early}$, which is different from the conventional VGG model. In addition, the later fusion function f_c is incorporated into the VGG architecture. Through the later fusion, two types of high-level features $\boldsymbol{v}$ and $f^{(i)}_{\boldsymbol{\theta}_{vgg}}$ are fused. This enables a spot-aware classification.

The proposed ROI-aware FES provides spot areas and encoded spot features, which are fed into the SAC-SubNet through early and later fusions to complete the entire network

to be trained in an end-to-end manner. Therefore, the ROI-aware FES is a novel spot feature extractor that encodes spot information, and subsequently teaches the SAC-SubNet regarding which areas and features should have a decisive role in classifying leaf diseases. The ROI-aware FES serves as a guide to achieve more accurate leaf disease identification. Because the spot features are utilized in SAC-SubNet, the leaf spot attention mechanism is reflected. Even though spatial and channel weight maps are not designed like other attention models in [22,42,43], our experiment revealed that simple concatenation is sufficient to develop a leaf spot attention mechanism and improve leaf disease identification.

The architecture of the ROI-aware FES in Figure 3 is inspired by semantic segmentation in [41]. Even though the proposed ROI-aware FES includes the leaf segmentation module, the goal of this study is different from that in [41]; in other words, our goal is not to divide the input image into multiple regions, but to encode spot features and achieve apple leaf disease identification. Naturally, the proposed architecture is different from those of TL-based methods [15–21] because two types of subnetworks are connected to create a whole network that is subsequently trained in an end-to-end manner. In other words, conventional TL-based methods do not include the ROI-aware FES. If the ROI-aware FES and ROI-aware feature fusion are removed from Figure 3, the proposed architecture becomes identical to the conventional VGG network. Therefore, whether recognition accuracy can be increased must be verified by comparing the performance between the proposed LSA-Net and the conventional VGG network [13].

2.2. Proposed Attention-Enhanced YOLO Network for Leaf Disease Detection

It is unclear whether the proposed ROI-aware FES can also be applied for leaf disease detection. Unlike leaf disease identification, leaf disease detection requires localization. To realize this, a new AE-YOLO model is proposed, as illustrated in Figure 4. It is well known that the existing YOLO model [33] consists of a feature extractor and detection subnetwork. However, this YOLO model excludes leaf spot attention mechanisms; thus, it lacks the ability to find spot areas that are spatially important and extract informative spot features. To complement this, the leaf spot attention mechanism needs to be incorporated into the YOLO network. As illustrated in Figure 4, the proposed AE-YOLO model contains ROI-aware FES and ROI-aware feature fusion.

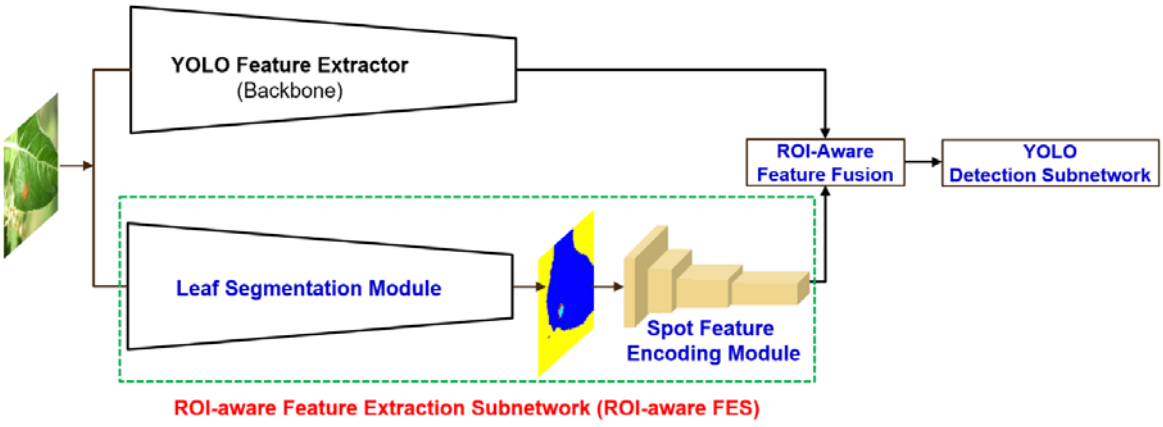

Figure 4. Proposed attention-enhanced YOLO network for apple leaf disease detection.

In the proposed AE-YOLO model, there are two branches for feature extraction, as illustrated in Figure 4. The first is the YOLO feature extractor, which is also called a backbone and is required for object detection models. Various image classification models, such as VGG [13] and ResNet [14], have been adopted for the backbone.

$$\boldsymbol{F}^{yolo} = f^{(i)}_{\boldsymbol{\theta}_{yolo}}(\boldsymbol{I}) \tag{5}$$

$f^{(i)}_{\boldsymbol{\theta}_{yolo}}$ denotes the ith layer of the YOLO feature extractor, and $\boldsymbol{F}^{yolo}$ indicates the output feature maps.

The second is ROI-aware FES. As illustrated in Figure 4, the ROI-aware FES is composed of two modules: a leaf segmentation module and a spot feature encoding module. Similar to the LSA-Net, the leaf segmentation module divides the input leaf image into background, leaf, and spot areas, and the spot feature encoding module transforms a high-dimensional segmented feature map into a low-dimensional feature vector to encode spot information and increase the representation power. The architecture of the leaf segmentation module is identical to that of LSA-Net. For simplicity, the spot feature encoding module has only convolution blocks. The output feature maps of the spot feature encoding module should have a size similar to $\boldsymbol{F}^{yolo}$.

$$\boldsymbol{F}^{roi} = f_{\theta_{roi}}(\boldsymbol{I}) \tag{6}$$

where $f_{\boldsymbol{\theta}_{roi}}$ indicates the ROI-aware FES, and $\boldsymbol{F}^{roi}$ denotes the output feature maps of $f_{\boldsymbol{\theta}_{roi}}$.

In the proposed AE-YOLO model, the ROI-aware feature fusion is also required to combine the two types of features: $\boldsymbol{F}^{yolo}$ and $\boldsymbol{F}^{roi}$. Unlike the LSA-Net that uses only a concatenation layer for ROI-aware feature fusion, our proposed AE-YOLO network adds more convolution blocks, as shown in Figure 5. To implement the spot-feature-guided attention model, the two features $\boldsymbol{F}^{yolo}$ and $\boldsymbol{F}^{roi}$ are first fused through the concatenation layer and subsequently passed through several convolution blocks.

$$\boldsymbol{F}^{a} = f_c\left(\boldsymbol{F}^{roi}, \boldsymbol{F}^{yolo}\right) \tag{7}$$

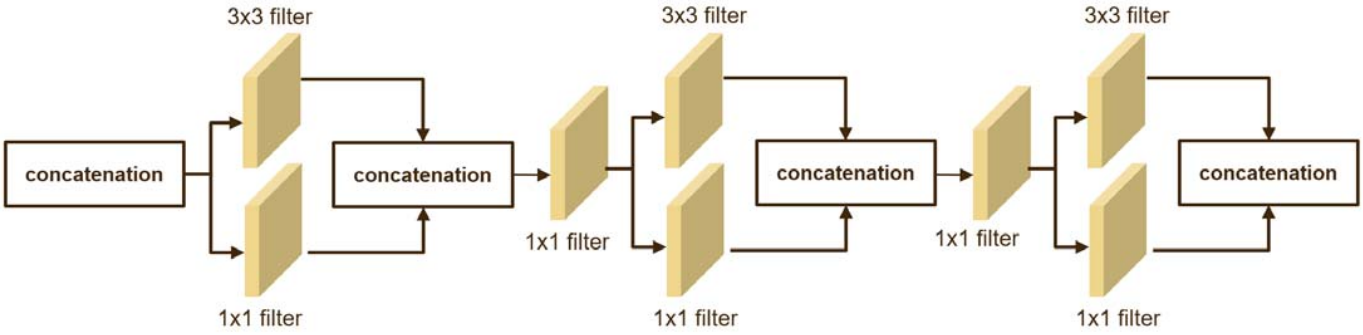

Figure 5. Proposed ROI-aware feature fusion for apple leaf disease detection.

Here, f_c is the function for the ROI-aware feature fusion, and $\boldsymbol{F}^{a}$ denotes the fused feature maps after applying the function f_c. In other words, $\boldsymbol{F}^{a}$ is the output of the last concatenation layer in Figure 5. This fused feature map $\boldsymbol{F}^{a}$ is fed into the YOLO detection subnetwork, as shown in Figure 4.

The architecture shown in Figure 5 is similar to the attention model used in SqueezeNet [43]. The only difference is that the proposed method uses two types of features, $\boldsymbol{F}^{yolo}$ and $\boldsymbol{F}^{roi}$, as inputs for more effective leaf spot attention modeling. In other words, the additional information on the spot features $\boldsymbol{F}^{roi}$ can serve as a guide to reinforce the leaf spot attention function. In Figure 5, the filter sizes used in the convolution layers are 1×1 and 3×3, which are used to model the channel and spatial attention, respectively. More details for channel and spatial attention layers are provided in Ref. [43].

In the proposed AE-YOLO model, there are two types of feature extractors. One is the YOLO feature extractor, and the other is the ROI-aware FES. The YOLO feature extractor can produce abundant features; however, it lacks the ability to determine spot areas. ROI-aware FES can provide more informative spot features. Therefore, the ROI-aware feature fusion can improve the discriminative power for better classification, which eventually

leads to an improvement in the object detection accuracy. The ROI-aware feature fusion can be regarded as a simple attention mechanism because it can teach the YOLO detection subnetwork which areas and features are more crucial for leaf disease detection. More details on the YOLO detection subnetwork and loss function are provided in [33].

3. Experiments

The proposed LSA-Net and AE-YOLO networks were implemented using MATLAB, and trained with four Titan-XP GPUs on a Windows operating system. Two experiments were conducted in this study. One was leaf disease identification, and the other was leaf disease detection. Leaf disease identification is to predict label values corresponding to leaf diseases from input leaf images, and leaf disease detection is to determine the bounding boxes surrounding diseased leaves. First, to compare the performance of the leaf disease identification, state-of-the-art classification models, that is, VGG [13], ResNet [14], and SqueezeNet [43], feature pyramid network (FPN) [23], attention gated network (AGN) [22], and pyramid vision transformer (PVT) [24] were tested. Second, to compare the performance of the leaf disease detection, state-of-the-art detection models such as YOLO [33], RCNN [30], Fast RCNN [31], RetinaNet [44], and Faster RCNN [32] were tested. For quantitative evaluation, the correct recognition accuracy (CRC) [40] and mean averaged precision (mAP) [30] were calculated. The dataset and source codes of the proposed LSA-Net and AE-YOLO can be downloaded at https://github.com/cvmllab/ (accessed on 27 August 2021).

3.1. Image Collection

All apple leaf images used in this study were provided by the Apple Research Institute in our country. The apple leaf images were categorized into three groups, according to two types of leaf diseases and normal leaf. Figure 6 shows examples of apple leaf images. The first row shows normal leaf images, and the second and third rows show diseased leaf images. Particularly, for the diseased leaf images with marssonina blotch, as shown in the second row, the blotch colors are similar to the normal leaf colors in the background. Therefore, the color-based clustering algorithm [39] might fail to extract the blotch colors from the normal leaves. This reveals that the leaf areas, background, and spot area must be divided. In addition, a real environment was considered to some degree, in that the leaf images were more complicated in the background than those tested in [17,21,39], where the background colors were nearly solid. In our database, the total numbers of normal leaf images, diseased leaf images with marssonina blotch, and diseased leaf images with alternaria leaf spot were 558, 2281, and 896, respectively.

3.2. Leaf Segmentation Module Training

Before training the entire network in an end-to-end manner, the leaf segmentation module was pretrained, as illustrated in Figures 3 and 4. Ground truth segmentation maps are required to train the leaf segmentation module. In this study, ground truth segmentation maps were generated manually through image editing to divide them into three areas: background, leaf area, and spot area. During image editing, the leaves without diseases were classified by background in the image, and those with diseases were classified by leaf area in the image. This simplifies the labeling process. Given the ground truth segmentation maps, the leaf segmentation module was trained via mini-batch gradient descent optimization [45].

Figure 6. Example of apple leaf images: normal leaf images (**first row**), diseased leaf images with marssonina blotch (**second row**), and diseased leaf images with alternaria leaf spot (**third row**).

3.3. Entire Network Training

The pretrained leaf segmentation module was adopted to train the LSA-Net and AE-YOLO networks. First, in the case of LSA-Net, the last loss layer of the leaf segmentation module is removed, and the spot feature encoding module is added to the back of the segmentation module, to construct the ROI-aware FES. Subsequently, the ROI-aware FES is connected with the SAC-SubNet through early and later ROI-aware feature fusions, thereby forming the entire network, as illustrated in Figure 3. Second, the proposed AE-YOLO network also requires a pretrained leaf segmentation module, as illustrated in Figure 4. Similar to the LSA-Net, the pretrained leaf segmentation module without the loss layer is added to the spot feature encoding module to form the ROI-aware FES. Subsequently, the entire network is constructed by incorporating the ROI-aware FES into the YOLO framework. The two entire networks were trained in an end-to-end manner. Therefore, the leaf segmentation module changes its pretrained parameters to be adopted for a new task. In other words, the predicted segmentation map is adjusted during the entire network training, for more accurate leaf disease identification and detection.

It is worthwhile to check the output feature map of the leaf segmentation module. Figure 7 compares the ground truth segmentation maps and the predicted segmentation feature maps. The predicted segmentation feature maps were extracted from the trained LSA-Net. In Figure 7, it is noted that the predicted segmentation feature maps still possess discriminative features for different areas, that is, background, leaf area, and spot area. In particular, it is observed that the spot areas in the leaves have blue colors, which are clearly distinguishable from the background's red colors, and the leaves' green colors. This indicates that the leaf segmentation module can provide discriminative features. There are a few segmentation errors in the predicted segmentation feature maps. However, the ultimate objective of this study is not accurate segmentation, but leaf disease identification and detection. This implies that the leaf segmentation module is sufficient to serve as a guide to achieve more accurate leaf disease identification and detection. For reference, the mean accuracy of the leaf segmentation module, defined as the ratio of correctly classified pixels to total pixels for each class, is approximately 86%, and the mean intersection over union (IoU), also known as the Jaccard similarity coefficient, is approximately 69% [46].

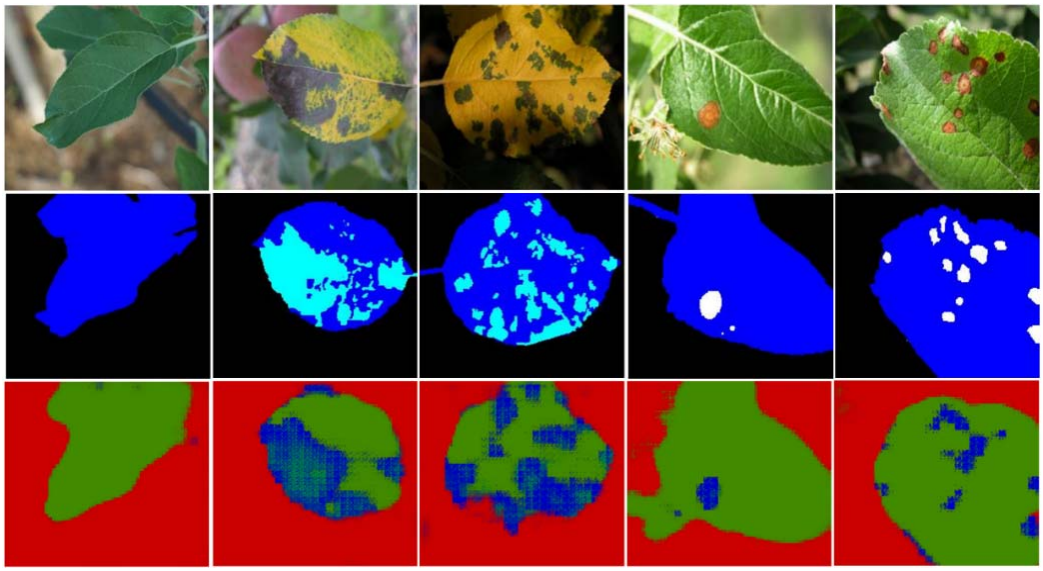

Figure 7. Leaf images (**first row**), ground truth segmentation maps (**second row**), predicted segmentation feature maps after training entire network (**last row**).

3.4. Performance Comparison

3.4.1. Leaf Disease Identification

Table 1 presents the CRC results for the proposed method and conventional state-of-the-art methods. In Table 1, all networks, including VGG, ResNet, FPN, AGN, SqueezeNet, and PVT, were initialized with the pretrained parameters with an ImageNet dataset, and TL was applied to each network with a new apple leaf dataset. Except for PVT, FPN, and AGN, library functions of MATLAB (2021a) were used to train the conventional classification models: VGG, ResNet, and SqueezeNet. The optimizer used was stochastic gradient descent (SGD) [45] with momentum. The epoch number was 30, and the batch size was set to 10. The learning rate was 0.001, and the momentum term was set to 0.9. The regularization term was ℓ_2-norm, and its weight was set to 0.0001. For more detailed parameter settings, please refer to the author's source code. For the PVT, the open-source code provided by the author in [24] was used with the default setting. AGN and FPN were implemented using MATLAB's layer functions. Therefore, the same training parameters were used for the conventional and proposed models, except for PVT.

Table 1. Performance evaluation for leaf disease identification.

Methods	CRC (%)
VGG16 [13]	90.19
ResNet50 [14]	87.87
Attention Gated Network (AGN) [22]	87.69
Feature Pyramid Network (FPN) [23]	88.58
SqueezeNet [43]	92.24
Pyramid Vision Transformer (PVT) [24]	93.70
Proposed LSA-Net	**96.07**

In Figure 3, if the ROI-aware FES and ROI-aware feature fusion are excluded from the LSA-Net, the proposed architecture becomes identical to the conventional VGG network. Therefore, it should be checked whether CRC can be improved with additional ROI-aware FES and ROI-aware feature fusion. In Table 1, by comparing the proposed LSA-Net and VGG16, it is known that the additional use of the ROI-aware FES and feature fusion

increased the CRC value by approximately 6%. This reveals that the ROI-aware FES and feature fusion can teach the SAC-SubNet regarding which areas and features should have a decisive role in classifying apple leaf diseases. The ROI-aware FES and feature fusion served as a guide for a more accurate leaf disease identification. In addition, the proposed LSA-Net demonstrated the best performance among all the methods.

Although VGG and ResNet can be adopted for the identification of leaf diseases, these models have limitations in improving the discriminative power because they do not model leaf spot attention mechanisms to extract discriminative and informative features from leaf images. In contrast, the SqueezeNet and PVT include attention mechanisms such as spatial and channel attention; thus, their CRC values are higher than those of VGG and ResNet. However, their performance is not better than that of the proposed LSA-Net because the SqueezeNet and PVT are self-attention vision models, which means that these networks do not utilize side information such as leaf segmentation and gradient maps. However, the proposed LSA-Net models the leaf spot attention mechanism based on the predicted leaf segmentation to extract spot features. This result confirms that the proposed leaf spot attention model is more effective than the self-attention vision model for leaf disease identification. The AGN is also a self-attention vision model; however, its performance is not better than that of SqueezeNet or PVT. Originally, the AGN was designed for medical image analysis, and it seems that the AGN is not suitable for leaf disease identification. In Table 1, the FPN adopted ResNet50 as the backbone. The main difference between FPN and ResNet50 is the pyramidal feature hierarchy, which has semantics from low to high levels. From Table 1, it is observed that the application of pyramidal feature hierarchy slightly increases the CRC value.

3.4.2. Leaf Disease Detection

Table 2 presents the mAP results for the proposed AE-YOLO and conventional object detection models. In Table 2, the conventional detection models, including the RCNN series and YOLO, selected ResNet50 as the backbone. Except for RetinaNet, library functions of MATLAB (2021a) were used with the same parameter settings to train the conventional object detection models. The optimizer used was SGD [45] with momentum. The epoch number was 100, and the batch size was 4. The learning rate was 0.001, and the momentum term was set to 0.9. The regularization term to prevent overfitting was ℓ_2-norm, and its weight was set to 0.0001. The data augmentation technique, including contrast, saturation, and brightness, was applied to increase the training data and prevent overfitting. For more detailed parameter settings, please refer to the author's source code. For RetinaNet, the default settings provided by the source code in [24] were used. The backbone used in RetinaNet was PVT.

Table 2. Performance evaluation for leaf disease detection.

Methods	AP (%) (Marssonina)	AP (%) (Alternaria)	mAP (%)
RCNN [30]	17.66	23.70	**20.70**
Fast RCNN [31]	46.10	38.00	**42.10**
Faster RCNN [32]	50.01	44.90	**47.50**
RetinaNet [44]	54.60	40.20	**47.40**
YOLO [33]	38.70	31.60	**35.10**
Proposed attention-enhanced YOLO (ResNet50)	54.30	47.40	**50.80**
Proposed attention-enhanced YOLO (VGG16)	55.00	42.40	**48.70**
Proposed attention-enhanced YOLO (SqueezeNet)	51.10	47.60	**49.40**

Similarly to the LSA-Net, it should be checked whether the use of the ROI-aware FES and ROI-aware feature fusion can lead to improvements in object detection accuracy. In Figure 4, if the ROI-aware FES and feature fusion are removed, the proposed AE-YOLO becomes identical to the conventional YOLO. By comparing the AP values of the proposed AE-YOLO and conventional YOLO in Table 2, it is known that the proposed AE-YOLO increases the AP value by 15.7% compared to the conventional YOLO. This result confirms that the ROI-aware FES and feature fusions have a significant effect on improving object detection performance.

In the last three rows of Table 2, the round bracket indicates the backbone adopted. In the proposed architecture illustrated in Figure 4, various feature extractors can be adopted for the backbone. In this study, three types of feature extractors, ResNet50, VGG16, and SqueezeNet, were tested. The last three rows indicate that the AP values depend on the backbone adopted, and ResNet is the best among the three feature extractors. This indicates that the feature extractors affect the final object detection performance. In the proposed AE-YOLO network, the ROI-aware FES is another feature extractor for spot detection. This ROI-aware FES can enhance the discriminative power of the YOLO feature extractor through ROI-aware feature fusion. This is the reason why the proposed AE-YOLO is superior to the conventional object detection models.

In Table 2, it is noted that Fast RCNN and Faster RCNN are not better than the proposed AE-YOLO. The Fast RCNN and Faster RCNN utilize region proposal, whereas the proposed AE-YOLO adopts an additional spot feature extractor, that is, ROI-aware FES. This result indicates that the ROI-aware FES is more effective in improving leaf disease detection than the region proposal. As shown in Table 2, the mAP of RetinaNet was 47.40, and its performance was comparable to that of the Faster RCNN. It is generally known that single-shot detectors are less accurate than two-stage detectors (e.g., Faster RCNN).

Figure 8 illustrates the predicted bounding boxes with the proposed AE-YOLO. As illustrated in the figures, the proposed AE-YOLO can detect diseased leaves properly. However, the diseased leaves were not discovered in the proposed network. In particular, when the leaves are small or covered by other leaves, and when the spot area is not clearly visible, false detection occurs. Since the leaves vary in size, and the occlusion problem is severe, leaf disease detection is more challenging than conventional object detections such as faces, cars, and pedestrians. Accordingly, the overall mAP values were low compared to those in [32].

Figure 8. Detection results; ground truth bounding boxes (**upper row**), predicted bounding boxes with the proposed AE-YOLO model (**bottom row**).

4. Conclusions

Novel leaf-spot attention networks for leaf disease identification and detection were introduced in this study. To improve the performance of leaf disease identification and detection, it is necessary to develop a novel ROI-aware feature extractor that can find spot areas from leaf images and encode semantic spot information. Based on the observation that leaf diseases exist in the leaf area, the ROI-aware feature extractor was designed to have two modules: leaf segmentation and spot feature encoding. First, the leaf segmentation module was pretrained to determine spot areas, then the spot feature encoding module was applied to extract informative spot features. Next, the ROI-aware feature extractor was connected to the ROI-aware feature fusion layer to model the leaf spot attention mechanism, and to be joined with the SAC-SubNet or YOLO detection subnetwork. During the entire network training, the ROI-aware feature extractor could teach the SAC-SubNet and YOLO detection subnetwork which areas and features should have a decisive role in classifying and localizing leaf diseases. The experimental results confirmed that the ROI-aware feature extractor and feature fusion can increase the performance of leaf disease identification and detection by boosting the discriminative power of spot features. It was also revealed that the proposed LSA-Net and AE-YOLO are superior to state-of-the-art deep learning models. In the future, we will test whether the proposed method can be extended to other applications such as pest detection and tomato leaf disease identification.

Funding: This work was carried out with the support of Cooperative Research Program for Agriculture Science & Technology Development (Grant No. PJ0163032021), National Institute of Crop Science (NICS), Rural Development Administration (RDA), Republic of Korea.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://github.com/cvmllab/ (accessed on 6 August 2021).

Conflicts of Interest: The author declares no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the Neural Information Processing Systems, Harrah's Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1–9.
2. Lowe, D.G. Distinct image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
3. Guo, Z.; Zhang, L.; Zhang, D. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.* **2010**, *19*, 1657–1663. [PubMed]
4. Yang, J.; Yu, K.; Gong, Y.; Huang, T. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1794–1801.
5. Singh, V.; Misra, A.K. Detection of plant leaf diseases using image segmentation and soft computing techniques. *Inf. Process. Agric.* **2017**, *4*, 41–49. [CrossRef]
6. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; pp. 2169–2178.
7. Perronnin, F.; Dance, C. Fisher Kernels on Visual Vocabularies for Image Categorization. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
8. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
9. Islam, M.; Dinh, A.; Wahid, K. Detection of potato diseases using image segmentation and multiclass support vector machine. In Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, Windsor, ON, Canada, 30 April–3 May 2017; pp. 1–4.
10. Quinlanm, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [CrossRef]
11. Chuanlei, Z.; Shanwen, Z.; Jucheng, Y.; Yancui, S.; Jia, C. Apple leaf disease identification using genetic algorithm and correlation based feature selection method. *Int. J. Agric. Biol. Eng.* **2017**, *1*, 74–83.
12. Aharon, M.; Elad, M.; Bruckstein, A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.* **2006**, *54*, 4311–4322. [CrossRef]

13. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
15. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. [CrossRef]
16. Fuentes, A.; Yoon, S.; Kim, S.C.; Park, D.S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* **2017**, *17*, 2022. [CrossRef]
17. Mohanty, S.P.; Hughes, D.; Salathe, M. Using deep learning for image-based plant disease detection. *Front. Plant Sci.* **2016**, *7*, 1419. [CrossRef]
18. Lu, Y.; Yi, S.; Zeng, N.; Liu, Y.; Zhang, Y. Identification of rice diseases using deep convolutional neural networks. *Neurocomputing* **2017**, *267*, 378–384. [CrossRef]
19. Tetila, E.C.; Machado, B.B.; Menezes, G.K.; Oliveira, A.; Alvares, M.; Amorim, W.P.; Belete, N.; Silva, G.; Pistori, H. Automatic recognition of soybean leaf diseases using UAV images and deep convolutional neural networks. *IEEE Geosci. Remote. Sens. Lett.* **2020**, *17*, 903–907. [CrossRef]
20. Bjerge, K.; Nielsen, J.B.; Sepstrup, M.V.; Helsing-Nielsen, F.; Hoye, T.T. An automated light trap to monitor moths (Lepidoptera) using computer vision-based tracking and deep learning. *Sensors* **2021**, *21*, 343. [CrossRef] [PubMed]
21. Wspanialy, P.; Moussa, M. A detection and severity estimation system for generic diseases of tomato greeenhouse plants. *Comput. Electron. Agric.* **2020**, *178*, 105701. [CrossRef]
22. Schlemper, J.; Oktay, O.; Schaap, M.; Heinrich, M.; Kainz, B.; Glocker, B.; Rueckert, D. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* **2019**, *53*, 197–207. [CrossRef]
23. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–25 July 2017; pp. 2117–2125.
24. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv* **2021**, arXiv:2102.12122.
25. Kumar, N.; Belhumeur, P.N.; Biswa, A.; Jacobs, D.W.; Kress, W.J.; Lopez, I.; Soares, J. Leafsnap: A Computer Vision System for Automatic Plant Species Identification. In Proceedings of the European Conference on Computer Vision, Florence Italy, 7–13 October 2012; pp. 502–516.
26. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
27. Uijlings, J.R.; van de Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
28. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef]
29. Zitnick, C.L.; Dollár, P. Edge Boxes: Locating Object Proposals from Edges. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 8–11 September 2014; pp. 391–405.
30. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
31. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
32. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only Look Once: Unified, Realtime Object Detection. In Proceedings of the IEEE Conference Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
34. Liu, W.; Anguelov, D.; Drhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
36. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
37. Jiang, E.; Chen, Y.; Liu, B.; He, D.; Liang, C. Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks. *IEEE Access* **2019**, *7*, 59069–59080. [CrossRef]
38. Tian, Y.; Yang, G.; Wang, Z.; Wang, H.; Li, E.; Liang, Z. Apple detection during different growth stages in orchards using improved YOLO-V3 model. *Comput. Electron. Agric.* **2019**, *157*, 417–426. [CrossRef]
39. Ali, H.; Lali, M.I.; Nawaz, M.Z.; Sharif, M.; Saleem, B.A. Symptom based automated detection of citrus diseases using color histogram and textural descriptors. *Comput. Electron. Agric.* **2017**, *138*, 92–104. [CrossRef]
40. Yu, H.-J.; Son, C.-H. Leaf Spot Attention Network for Apple Leaf Disease Identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 52–53.

41. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the International Conference on Neural Information Processing System, Long Beach, CA, USA, 4–9 December 2017.
43. Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5MB model size. *arXiv* **2016**, arXiv:1602.07360.
44. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
45. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
46. Csuka, G.; Larlus, D.; Perronnin, F. What Is a Good Evaluation Measure for Semantic Segmentation? In Proceedings of the British Machine Vision Conference, Bristol, UK, 9–13 September 2013; pp. 32.1–32.11.