Computers and Electronics in Agriculture 195 (2022) 106853

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

An efficient attention module for instance segmentation network in pest monitoring

Hanxiang Wang [a,1], Yanfen Li [a,1], L. Minh Dang [b], Hyeonjoon Moon [a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea
[b] Department of Information Technology, FPT University, Ho Chi Minh city, Viet Nam

ARTICLE INFO

Keywords:
Crop pest
Image processing
Deep learning
Instance segmentation
Attention mechanism

ABSTRACT

Prompt and specialized pest management involving localization and recognition has become a crucial means to prevent pest attacks in modern agriculture. Traditional pest monitoring methods are inaccurate and inefficient due to the hand-crafted features and the low-resolution images. As a result, this study presents an automatic framework that can precisely detect 10 species of pests in the natural environment and assist humans in identifying the locations and contours of pests efficiently. The main contributions of this paper include (1) a novel attention module that encourages the network to focus on the important features; (2) an optimized super-resolution approach that is used for both training and testing images to enhance the image quality; (3) a pest monitoring network is proposed by improving the D2Det's structure and adjusting the parameters; and (4) a dataset containing pest images and manually annotated files. Experiments showed that the proposed Pest-D2Det model achieved state-of-the-art performances in terms of the mean Average Precision (mAP) values of detection (78.6%) and segmentation (77.2%). Meanwhile, the performance of our efficient channel and spatial attention network (ECSA-Net) indicated that it is lightweight and effective, which can be integrated into deep learning-based models without computation burden.

1. Introduction

Crops are damaged by various species of pests every year during the cultivation process, which makes the yield and quality of crops decline in varying degrees. Prompt management of pests can effectively avoid the economic loss of crops. However, pests have the characteristics of different scales, variant shapes, and complex textures. Traditional methods that only rely on vision and experience cannot identify and localize a large number of pests quickly and accurately under complex environments. In modern agriculture, the urgent demand for pest monitoring and control has driven the further exploration of intelligent pest localization and recognition systems.

Previous studies of pest localization are mainly based on image processing and conventional machine learning approaches due to simple implementations (Maharlooei et al., 2017; Wang et al., 2018). Nevertheless, these methods are easy to be affected by the change of image features. Recently, deep learning methods are employed in pest instance identification tasks on account of excellent performances (Chen et al., 2021; Liu, 2019). In particular, the attention mechanism was proved to be active in enhancing the perception of the target area (Zhao et al., 2021). In this context, this research aims at designing a novel attention mechanism and integrating it into a pest segmentation network to boost overall accuracy.

Moreover, as long as the collected training data are diverse and clear enough, the existing technologies are able to handle the pest localization problems. However, the captured images from natural scenes vary at resolutions and illumination conditions in practice, which is challenging for the multi-class detection and identification task (Barbedo, 2020)(Li, 2022). Therefore, super-resolution (SR) and data augmentation techniques are explored before the pest localization network in order to provide more satisfying image qualities in this study.

The principal contributions of this study are listed as follows.

- A novel attention module is designed to encourage the network to focus on the informative features.
- An optimized SR method is used to improve the performances for both detection and segmentation tasks.

* Corresponding author.
E-mail addresses: hanxiang@sju.ac.kr (H. Wang), 1826535091@sju.ac.kr (Y. Li), hmoon@sejong.ac.kr (H. Moon).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.compag.2022.106853
Received 6 July 2021; Received in revised form 13 December 2021; Accepted 4 March 2022

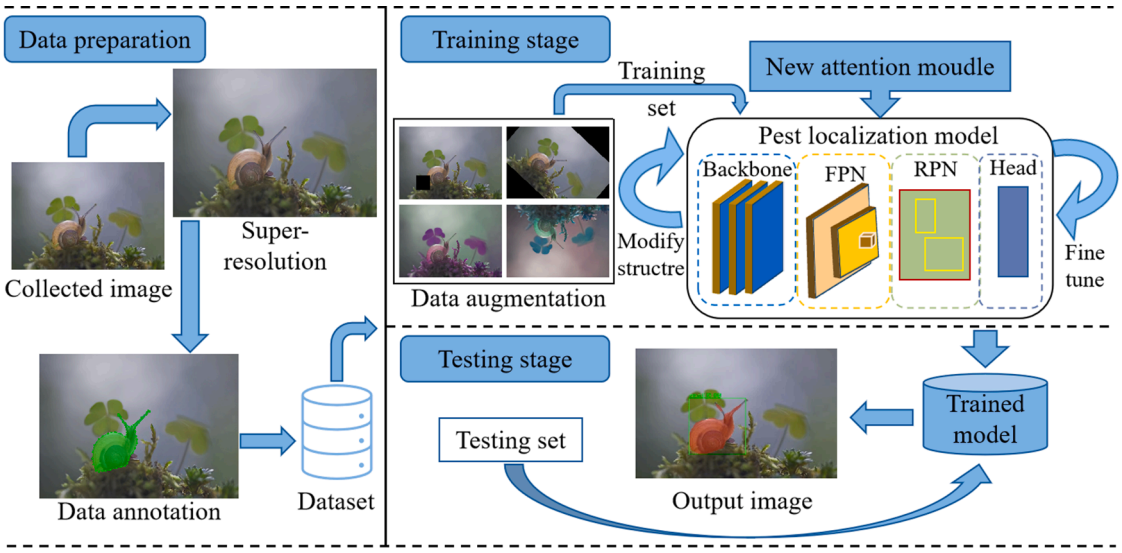

Fig. 1. Diagram of the proposed pest localization system. In the data preparation stage, the dataset consists of the preprocessed images and the corresponding annotation files. The pest localization model is proposed by modifying the structure and tuning the parameters in the training stage. And then the performance of the trained model is evaluated in the testing stage.

- An effective pest monitoring network is proposed by modifying the D2Det's structure and adjusting the parameters.
- A dataset including 9,872 pest images and corresponding annotated files is provided, which can be used in the instance segmentation task.

This paper contains six sections. Related studies on pest analysis are reviewed in Section 2. Section 3 presents the overall flowchart of the present framework and the primary methodology. Section 4 introduces the dataset and corresponding metrics for evaluations. Some experiments are presented and discussed in Section 5. Finally, the conclusion of this study is given in Section 6.

2. Related work

Some early reports on pest location and surveillance explored traditional image processing and machine learning techniques. For example, some image processing methods, like color transformation and contrast adjustment, were employed to improve the efficiency of pest detection and counting on leaves (Maharlooei et al., 2017). Experiments suggested the image processing method for segmentation obtained good results in certain illumination conditions. In another work, a cognitive vision approach was used for pest segmentation by using some strategies such as block processing, adaptive initial cluster centers, and leaf vein removal (Wang et al., 2018). The results demonstrated the cognitive segmentation method with some complicated processing had a lower error rate and standard deviation. As one of the classical machine learning algorithms, a Support Vector Machine (SVM) was applied to classify and detect various types of pests. Before that, image processing was used to improve detection precision and speed (Ebrahimi et al., 2017). Besides, the SVM was also used and integrated into the study of bio-inspired techniques for pest image recognition. The model inspired by the human attention mechanism was used to detect pest areas, and the features are extracted by using a proposed method. After that, all features were input to an SVM model for the classification task. Experiments showed the developed method reached a result comparable to the deep learning methods (Deng et al., 2018).

Recently, deep learning techniques have been particularly trendy due to the ability of automatic feature extraction. A smartphone-based pest detection application was developed by employing different object detection models. Among the experimental models, YOLOv4 obtained the highest accuracy on the data with three classes (Chen et al., 2021). In (Liu, 2019), a PestNet model was introduced to detect and classify 16 species of pests based on fixed regions. A Channel-Spatial Attention was united into the backbone to extract features firstly, and then a potential pest location was obtained. Finally, the Position-Sensitive Score Map was used to get results of classification and detection. This study achieved a high mean Average Precision (mAP) of 75.46% for pest detection. Based on deep learning, a residual network with an attention mechanism was presented to detect the severity of the plant disease. Compared with other existing attention networks, the proposed model achieved a better classification accuracy and shorter speed (Zhao et al., 2021).

Deep learning-based pest monitoring frameworks need quality and sufficient data to support reliable detection and recognition results. Thus, some researchers attempted to tackle the data problems by using different image processing technologies. For instance, a data augmentation approach in (Li, 2019) is introduced to solve the problem that the collected images have significant differences in scale and posture. In another work, the captured images under the natural environment are noisy and blurry, which is not conducive to detect and segment pests in images. An SR method was used to boost the detection precision of low-resolution images (Yue et al., 2018). Besides, different image processing technologies were combined to improve the performance of the classification task in greenhouses (Espinoza et al., 2016) or natural scenes (Li et al., 2020).

Prior to our work, there is still a lack of a benchmark dataset for pest segmentation tasks. In addition, the conventional methods combining image processing and machine learning are easy to implement and have low computational complexity, but they rely heavily on handcrafted features. As a particular type of machine learning, deep learning methods require an extensive dataset and high computing resources. However, they can automatically extract features and get satisfactory results in computer vision. Based on the aforementioned problems, data augmentation approaches and the sparse convolution-based SR model are employed to build a dataset with more high-resolution images. After that, a pest localization network with an attention mechanism module is proposed to detect and classify pests in natural environments efficiently.

3. Proposed pest localization framework

In this section, the system overview of pest localization and

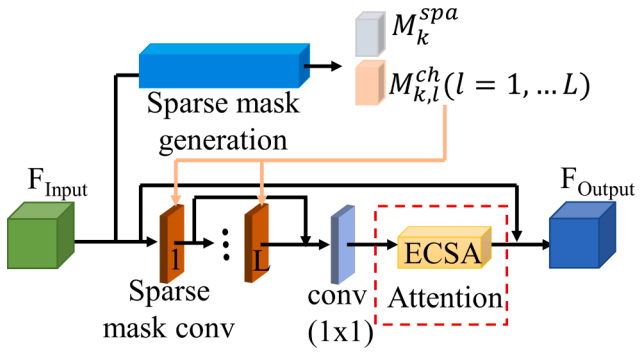

Fig. 2. The structure of the improved sparse mask module (SMM). **Note.** An efficient ECSA-Net in the red dashed box was proposed to replace the previous attention module, which reduces the parameters and computations of the original model. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

recognition is introduced in Fig. 1. Firstly, the collected images are preprocessed by applying the SR algorithm (Section 3.1). A segmentation dataset is then constructed by annotating the images, which includes the training set and testing set. In the training stage, 9,472 images are generated to enhance the generalization ability of the deep learning model by adopting some data augmentation methods, such as random rotations, flips, and brightness adjustment. After that, the training set is fed into a deep learning-based pest localization model. Various strategies are implemented including architecture modification, attention module integration (Section 3.2), and fine-tuning (Section 3.3) to improve the performance of the model. Finally, the effectiveness of the trained model is assessed by assessing the predicted results in the testing stage from the aspects of detection mAP and segmentation mAP, respectively.

3.1. Image preprocessing

The SR methods are usually carried out as a preprocessing method to enhance the image resolutions because the blurry tiny pest images captured in the natural environment are challenging for the multi-class detection and recognition tasks. In this study, the latest and efficient method named a Sparse Mask Super-resolution (SMSR) is explored to reconstruct pest images with detailed information (Wang, 2021). The novelty of the SMSR network is that the spatial and channel masks are used to distinguish different regions. The former can identify essential features, and the latter is used to mark unimportant features. Hence, the unimportant information in feature maps is skipped to release the computing space while keeping the equivalent performance.

The sparse mask module (SMM) is a crucial component of the SMSR network, which is mainly used to reduce redundancy by localizing and extracting important features. As shown in Fig. 2, spatial masks (M_k^{spa}) and channel masks ($M_{k,l}^{ch}$) are generated to obtain the redundant calculation of the model. Then, masks are sent to the corresponding sparse mask convolutional layers, ranging from 1 to L. After that, feature maps go through a 1x1 convolution layer and an attention module, where the attention module is used to encourage feature maps to be more informative and meaningful. After applying the proposed attention module, the parameters and computations of the original SMSR model were reduced.

In order to further reduce the parameters and computations of the SMSR model, the attention module named SENet (Hu et al., 2018) in SMM was replaced with our proposed attention module highlighted in the red dashed box. The model structure and complexity of our attention module will be introduced in Section 3.2. In the SR stage, a new dataset containing 1300 training samples and 150 testing samples was established to improve the model's generalization ability by combining the DIV2K dataset (Agustsson and Timofte, 2017) with our pest dataset.

3.2. Attention module: ECSA-Net

A convolutional block attention module named efficient channel and spatial attention network (ECSA-Net) is designed to recalibrate features from the aspects of channel and space. Fig. 3 shows the model structure of ECSA-Net, which includes the channel attention branch and spatial attention branch. In this study, the proposed lightweight ECSA-Net is integrated into the super-resolution model and the pest localization model without computational burden.

Inspired by the idea of feature aggregation in the convolutional block attention module (CBAM) (Woo et al., 2018), the average pooling and the max-pooling are applied in the channel attention branch to compress the dimension of the input feature F(HxWxC) along the channel direction. CBAM utilized two fully connected layers to reduce the dimensions of feature maps, which affects the learning effectiveness of the channel attention branch. Besides, adaptive convolution has a positive impact on improving the cross-channel learning ability (Wang et al., 2020). As a result, we applied a one-dimensional convolution layer that adaptively selects the convolution kernel size according to the number of channels after the pooling operation. The mapping relationship between the

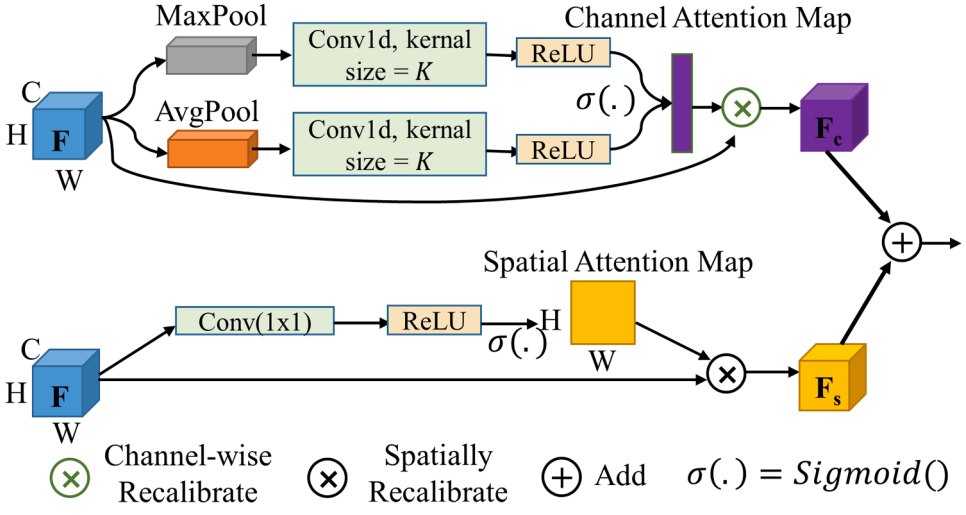

Fig. 3. The structure of the proposed ECSA-Net, which includes the channel attention branch and the spatial attention branch.

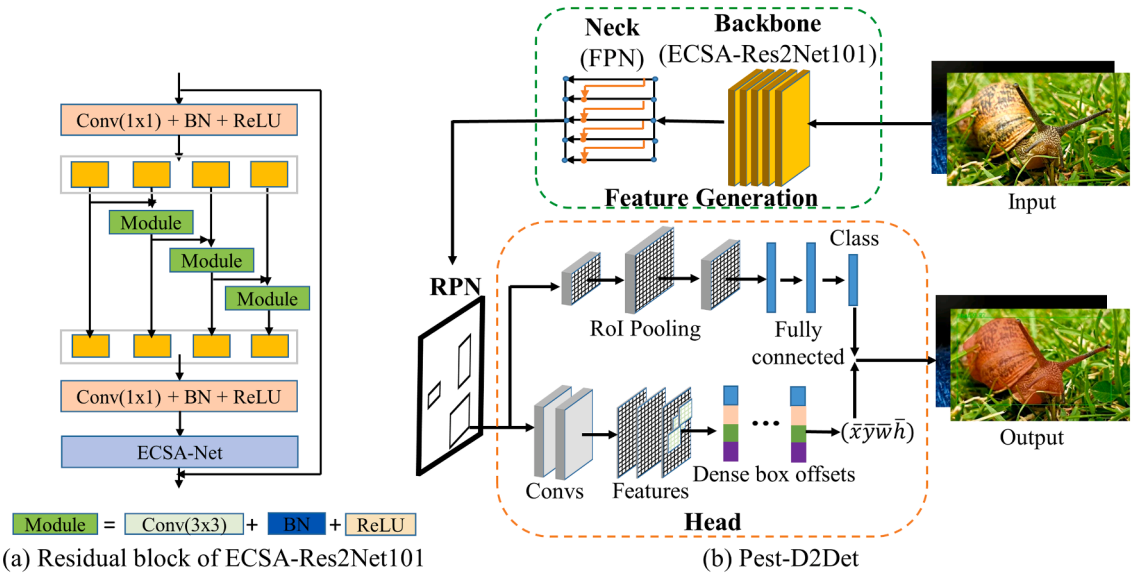

Fig. 4. (a) The residual block of the new backbone (ECSA-Res2Net101); (b) The overall structure of pest localization model (Pest-D2Det). **Note.** Images are fed into the proposed backbone and FPN module to generate features, RPN is used to classify the background and foreground of inputs. After that, there are 2 branches to predict the object area and label.

number of channels and the size of the convolution kernel is shown in Equation (2). When the number of channels increases, a larger convolution kernel is selected to control the range of cross-channel learning. This convolution layer with shared weights avoids the mismatch between channel and weight, and it dramatically reduces the parameter quantity. After the convolution operation, the feature maps on the two branches are combined together to form the channel attention maps through the ReLU activation function and the Sigmoid activation function. The feature map F_c is obtained by multiplying the channel attention map and the input feature map F, as shown in Equation (1).

In the spatial branch, the input feature F passes through a 2-dimensional convolution layer with a 1x1 kernel to generate a spatial projection tensor with the size of HxW. Each point P_{xy} on the projection tensor represents the combination of all channels at (x, y) position. After that, the ReLU layer and Sigmoid layer activate tensors in the range of [0, 1] to generate spatial attention maps. Equation (3) shows the calculation for the spatial feature map F_s. Finally, the spatial attention branch and the channel attention branch are connected in parallel, and the feature maps of the two branches are fused to generate a new feature map.

$$F_c = F^{*}\left(\sigma\left(r\left(f^{K}(Avg(F))\right)\right) + \sigma\left(r\left(f^{k}(Max(F))\right)\right)\right) \tag{1}$$

$$K = \frac{\log_2 C}{\alpha} + \beta/\alpha \tag{2}$$

$$F_s = F^{*}\sigma(r(f^{3\times 3}(F))) \tag{3}$$

where F_c and F_s represent output feature maps of the channel attention branch and the spatial attention branch, respectively. F refers to the input features of ECSA-Net, σ and r are Sigmoid activation function and ReLU activation function, respectively. The *Avg* function represents the average pooling, and *Max* function represents the max pooling. Convolution operation is represented by f, K and 3x3 indicate the kernel sizes of one-dimensional convolution and two-dimensional convolution, respectively. In Equation (2), the convolution kernel K has a nonlinear relationship with the number of channels C, constant α and β were set to 2 and 1 in this study. In convolution operation, the convolution kernel must be set to odd, so when K is even, K becomes $K + 1$.

3.3. Pest-D2Det's architecture

In this study, an effective Pest-D2Det model is presented to detect and identify pests by improving a two-stage D2Det model (Cao et al., 2020). As shown in Fig. 4 (b), the architecture of the Pest-D2Det model consists of backbone, neck, region proposal network (RPN) (He et al., 2017), and head. Firstly, Res2Net101 (Gao et al., 2019) replaces the previous backbone (ResNet101) in the D2Det model because of its powerful multi-scale feature representation ability. Then, a proposed lightweight attention module is embedded into each residual block of the Res2Net101 model to form a new backbone named ECSA-Res2Net101, as shown in Fig. 4 (a). Next, ECSA-Res2Net101 and feature pyramid networks (FPN) (Lin et al., 2017) are used to extract and fuse effective multi-level features from input images. Then, RPN generates the foreground candidate box by sliding the window on the feature map, and it calculates the position offset of the foreground candidate box relative to the Ground Truth (GT) (He et al., 2017). The head section of D2Det is retained in our model structure. The two branches of the head adopt the dense local regression and the discriminative ROI pooling algorithm (Cao et al., 2020) respectively to locate the target area and predict the target label.

The proposed ECSA-Net in the new backbone extracts more effective features related to spatial location and channel information for the backbone network, and it does not add a lot of parameters to the model. According to the experimental result, the parameters amount of the Pest-D2Det network is about 109.05 M, and the parameters of the ECSA-Net module (0.68 M) only account for 0.62% of the total parameters. On the other hand, the ECSA-Net module only occupies 0.78% of the total computations in terms of floating point operations per second (FLOPs).

In the training process, the loss function, optimizer, and initial learning rete of the model were fine-tuned to further improve the model's learning ability. Firstly, the Pest-D2Det network adopted different loss functions for regression and classification tasks. Focal loss is an improvement on cross-entropy loss, making the model pay more attention to the training samples with poor classification results by reducing the weight of samples that are easy to classify. Different from the classification task, the loss function of the regression task is used to calculate the difference between the predicted box and the GT. The CIoU loss function converges fast, and it involves the calculation of the overlap area, center point distance, and aspect ratios of predictions and

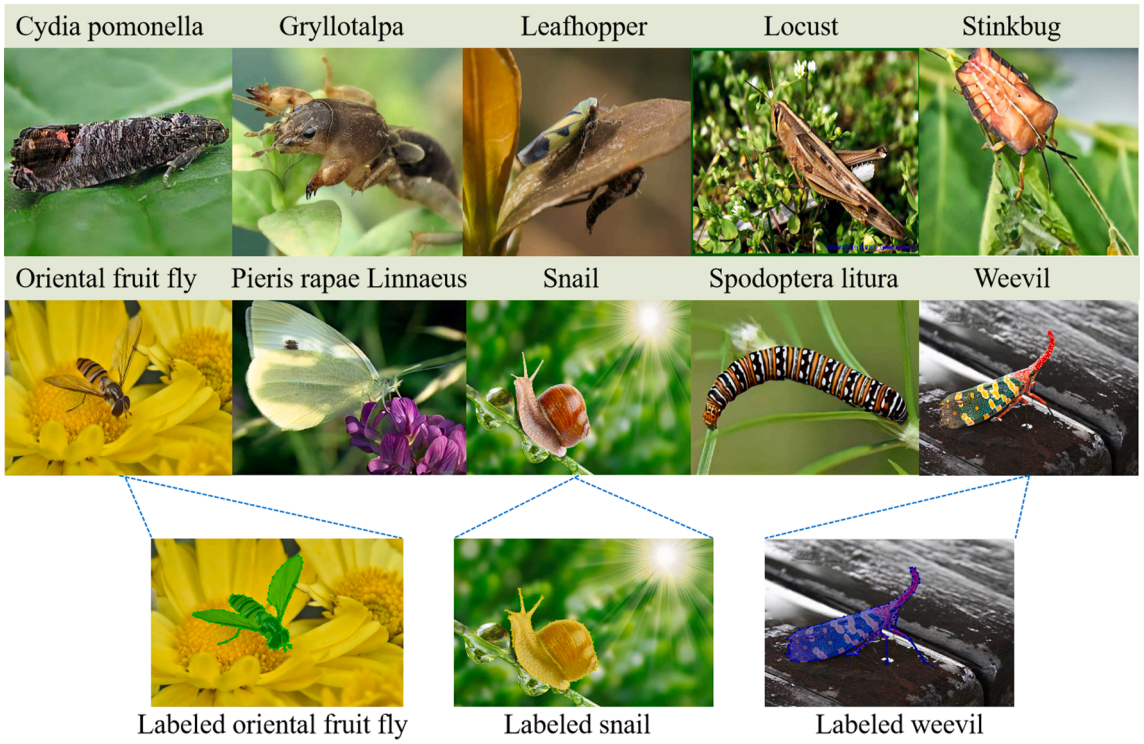

Fig. 5. Visualization of the proposed pest dataset, which includes the original images and the annotated images.

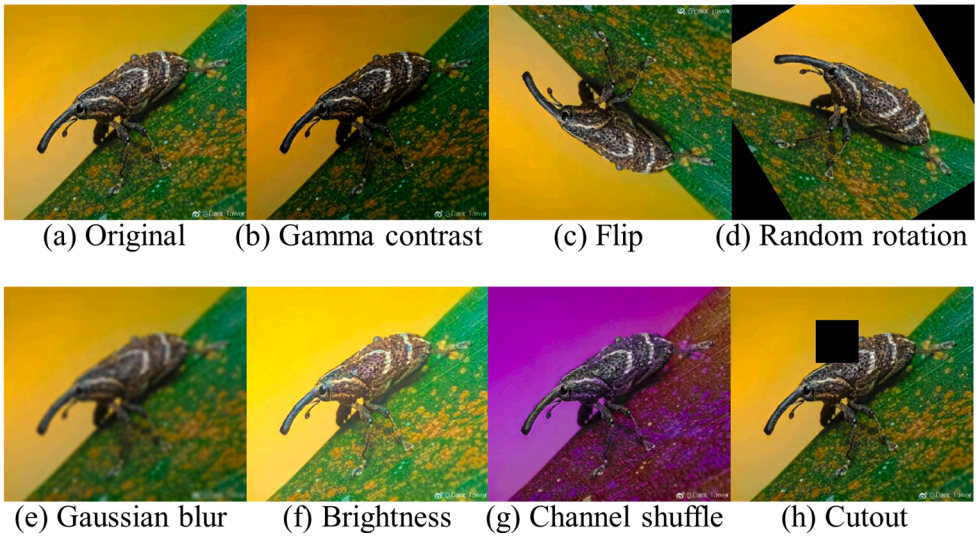

Fig. 6. Example of various data augmentation methods (a) original image, (b) Gamma contrast, (c) flip, (d) random rotation, (e) Gaussian blur, (f) brightness, (g) channel shuffle, and (h) cutout.

GTs. During the fine-tuning process, the influence of cross-entropy loss and focal loss for the classification task, as well as the effectiveness of CIoU loss and smooth L1 loss were analyzed and discussed in detail. Secondly, a more suitable optimizer was used to replace the SGD optimizer that was used in the D2Det model. As for the initial learning rate, it is usually set to 0.001, 0.01, or 0.1 in previous work (Kim et al., 2016) (Dang, 2022). Although high learning rate can boost the training convergence rate, it is prone to gradient exploding or vanishing. Therefore, the initial learning rate was finally adjusted from 0.001 to 0.01 to improve the training performance in terms of convergence speed and stability.

4. Dataset and evaluation metric

4.1. Dataset

4.1.1. Data preparation

Based on the crop pest dataset proposed in (Li et al., 2020), a total of 2,000 original images were manually reviewed and selected to construct a representative dataset for the pest instance localization and classification tasks. There are 10 types of pests, including cydia pomonella, gryllotalpa, leafhopper, locust, oriental fruit fly, Pieris rapae Linnaeus, snail, spodoptera litura, stinkbug, weevil. The pest images have variant resolutions, ranging from 224 × 90 to 4000 × 2337, and most of them

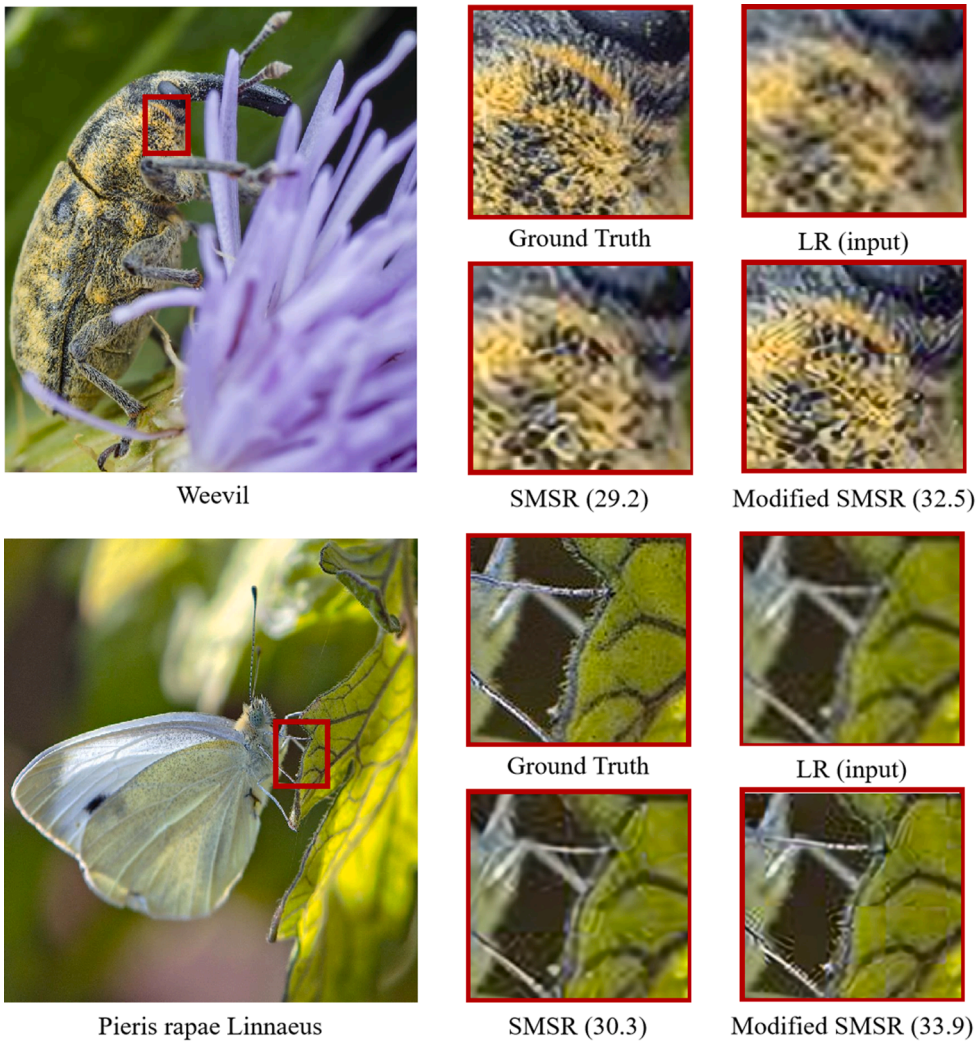

Fig. 7. Results of the original SMSR model and modified SMSR model with the scale factor x4. **Note.** 'LR' represents the low-resolution image that is used as the input data of the SR method. In the SR phase, the input LR images is processed to generate the output SR image, and then the obtained SR image is compared with the Ground Truth in order to evaluate the model's image reconstruction effect.

were acquired in natural scenes.

Before the pest segmentation process, all the pest images should be labeled with a public available annotation tool (LabelMe) to obtain the GT. The generated format of GT is a json file, and annotated pixels in GT indicate the pest location. After labelling, the images with annotation files were separated into a training set and validation set with a ratio of 8:2. The samples from the dataset for some original images and visualization annotations are shown in Fig. 5.

4.1.2. Data augmentation

Deep learning-based approaches require a sufficient number of samples. The more the number of training samples, the better the effect of the trained model and the stronger the generalization ability of the model. Thus, data augmentation was employed on all original training images as well as the labeled files to expand the data. After the data augmentation process, the total number of training images (9,472) is about five times more than the previous training amount (1,600).

Some proper augmentation techniques were applied to this study as an addition to the common data augmentation methods. For example, cutout randomly fills a fixed size square area with the specified pixels in an image, which enables the feature extractor to learn the global information of the image (DeVries and Taylor, 2017). Besides, the channel shuffle generates new data by randomly transforming the color space of the image to enhance the model's generalization ability in the training process (Jung, 2019). Fig. 6 illustrates some examples of various data augmentation methods.

4.2. Evaluation metric

This study mainly focuses on the performance of pest detection and segmentation tasks, and the corresponding metric suggested by some researchers in (Araujo, 2019; Hong, 2020) is mean Average Precision (mAP). This metric is calculated by the values of true positive (TP) and false positive (FP) to assess the detection and segmentation results. The calculation is shown in Equation (4).

$$mAP = mean(\frac{TP}{FP + TP}) \tag{4}$$

where TP means the number of target pixels that are accurately detected and segmented as target pixels. FP indicates the number of non-target pixels that are incorrectly detected and segmented as target pixels.

Moreover, peak signal to noise ratio (PSNR) is used to assess the effectiveness of the SR approach, which is an important metric in image-based instance localization and recognition (Sara et al., 2019). The PSNR value was computed using Equation (5) shown below:

Table 1
The performances of the original and the modified SMSR models on pest detection and segmentation.

Index	Model	Training set	Testing set	Det_AP				Seg_AP			
				PRL	WEL	CP	GLP	PRL	WEL	CP	GLP
1	Original SMSR	√	√	71.7	64.0	75.3	71.4	68.9	59.3	76.4	64.6
2	Modified SMSR	√	√	**73.4**	**64.5**	**76.8**	**72.5**	**69.7**	**61.2**	**78.9**	**66.3**
3		×	√	69.1	62.6	73.2	70.0	65.7	57.8	72.4	61.8
4		×	×	64.5	57.2	70.0	64.9	61.2	52.8	68.1	57.5
5		√	×	67.2	59.9	71.8	67.7	63.9	55.3	70.6	60.1

Note. 'Det_AP' and 'Seg_AP' represent the AP values for pest detection and segmentation, respectively. '√' indicates the images were processed by SR, and '×' indicates the images were not processed by SR. 'PRL' is the Pieris rapae Linnaeus class, 'WEL' is the weevil class, 'CP' is the cydia pomonella class, 'GLP' is the Gryllotalpa class.

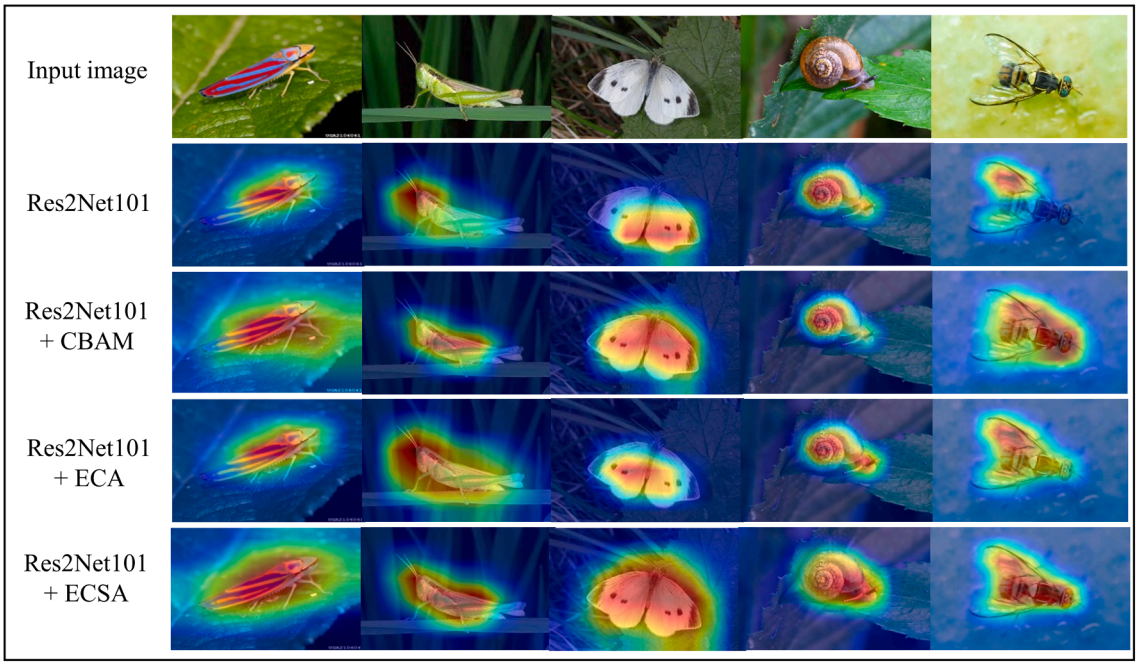

Fig. 8. Visualization results of the networks in the feature extraction process. The networks are integrated with and without different attention modules.

$$PSNR = 10\log_{10}\frac{m^2}{MSE} \tag{5}$$

where m is the maximum value of the input pixels, and the MSE represents the mean square error. In this study, a big PSNR value implies that there is a slight difference between the GT image and the reconstructed image.

In terms of the efficiency, the metric termed FLOPs is considered as a prevalent indicator to measure the model's computations (Molchanov et al., 2016). The calculation of the deep neural network is reflected in the convolution layer ($conv$) and the fully connected layer (fc), as shown in Equation (6) and (7).

$$FLOPs(conv) = 2{*}HW{*}(C_iK^2 + 1)C_o \tag{6}$$

$$FLOPs(fc) = 2{*}C_i{*}C_o \tag{7}$$

where H and W represent the height and width of the feature map after convolution operation, respectively. C_i and C_o is the number of input channels and output channels. K represents the size of the convolution kernel.

5. Experimental results

All the experiments were conducted on a Linux machine with the Ubuntu 18.04 system installed. It has four Tesla V100PCIe32GB GPUs, an Intel® Xeon® E5-2698 processor, and 256 GB of DDR4 RAM. In the training process of the proposed model, the focal loss was used for classification, and the CIoU loss was used for regression. The Adam optimizer is applied to update the training parameters. The initial learning rate was 0.01, and the decay was 0.0001. The batch size was set to 2, and the number of training epochs was 36. First of all, Section 5.1 demonstrates the effectiveness of the optimized SMSR network in the presented pest localization framework. Section 5.2 then showed how the proposed Pest-D2Det model performed on the proposed dataset. After that, Section 5.3 examines the improvement of the fine-tuning process. Next, the performances in various scenarios are discussed and analyzed in Section 5.4. Finally, the proposed instance segmentation model is compared with other work in Section 5.5.

5.1. Image preprocessing

In the image preprocessing phase, the super-resolution technique is performed to boost the resolution of the input pest image, and the performances of the SMSR model before and after modification are evaluated by calculating PSNR. In this experiment, two types of pests (weevil and Pieris rapae Linnaeus) were randomly selected to show the improvement of the modification on the original SMSR model. For each pest image, the regions of interest were cropped from the GT, low-resolution (LR) input image, result of the original SMSR model, result of modified SMSR model, respectively. The super-resolution results of two different models with the scale factor x4 are presented in Fig. 7, which suggested that the modified SMSR network successfully

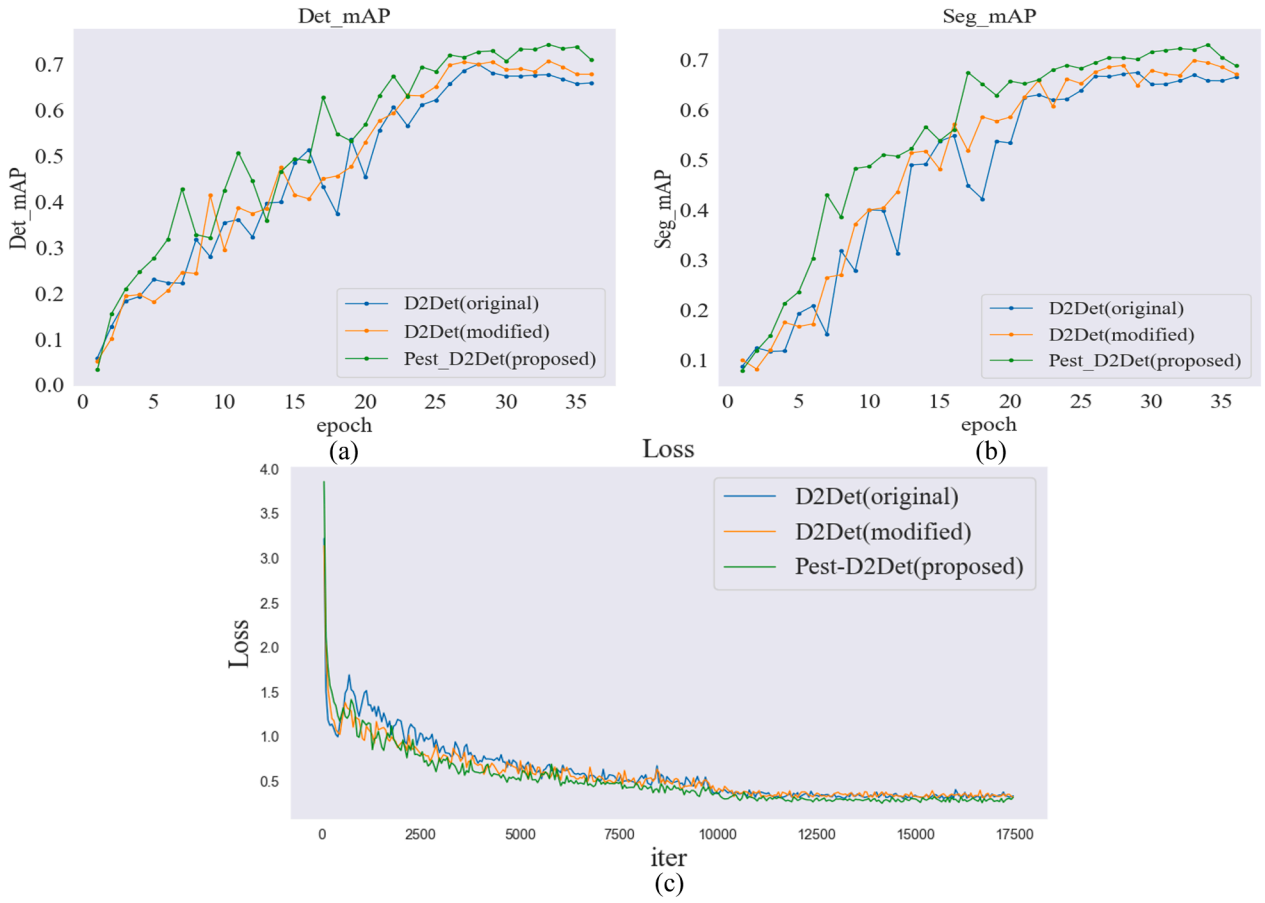

Fig. 9. (a) mAP curves of the detection task on the validation set, (b) mAP curves of the segmentation task on the validation set, and (c) loss function curves on the training set. **Note.** 'Det_mAP' and 'Seg_mAP' represent the mAP for pest detection and segmentation, respectively.

Table 2
Pest detection and segmentation results of different models. Comparison of different models from the aspects of model parameters (Param.), FLOPs, detection mAP, and segmentation mAP. mAP (mAP at IoU = 0.50:0.95). mAP50 (mAP at IoU = 0.50). mAP75 (mAP at IoU = 0.75).

Backbone	Framework	Param.	FLOPs	Detection			Segmentation		
				mAP	mAP50	mAP75	mAP	mAP50	mAP75
Res2Net101	D2Det	**108.37 M**	**341.11G**	70.6	89.7	86.1	69.6	88.7	85.4
Res2Net101 + SE		113.77 M	344.31G	71.9	91.5	87.2	71.0	90.1	87.2
Res2Net101 + CBAM		113.76 M	345.44G	72.4	92.1	88.5	71.4	90.9	87.9
Res2Net101 + ECA		109.05 M	342.12G	72.1	91.9	88.3	71.3	90.7	87.9
Res2Net101 + ECSA (proposed)		109.05 M	343.80G	**73.3**	**92.4**	**89.6**	**72.6**	**91.5**	**88.3**

Table 3
Different parameter settings (loss function, optimizer, and initial learning rete) and the corresponding performances in terms of detection mAP (Det-mAP) and segmentation mAP (Seg_mAP).

Fine-tuning	Loss function		Optimizer	Initial learning rate	Det_mAP	Seg_mAP
	Classification	Regression				
1	cross-entropy loss	smooth L1 loss	SGD	0.001	73.3	72.6
2	focal loss	CIoU loss	SGD	0.001	75.4	74.8
3	focal loss	CIoU loss	SGD	0.01	76.2	75.6
4	focal loss	CIoU loss	Adam	0.01	**78.6**	**77.2**

reconstructed the details from the seriously degraded inputs. For example, the texture of leaves and the contours of pests are blurry in the input images and the output images of the original SMSR, whereas these details are vivid in the results of the modified model. Besides, the modified SMSR model obtained better performance in terms of the PSNR values.

In this section, four classes (Pieris rapae Linnaeus, weevil, cydia pomonella, and Gryllotalpa) with inferior image quality were explored to highlight the significant impact of the SR phase on the whole framework. In addition, the effects of the SR model before and after modification are compared by evaluating their respective performances. Original images and SR images are used for training and testing to make

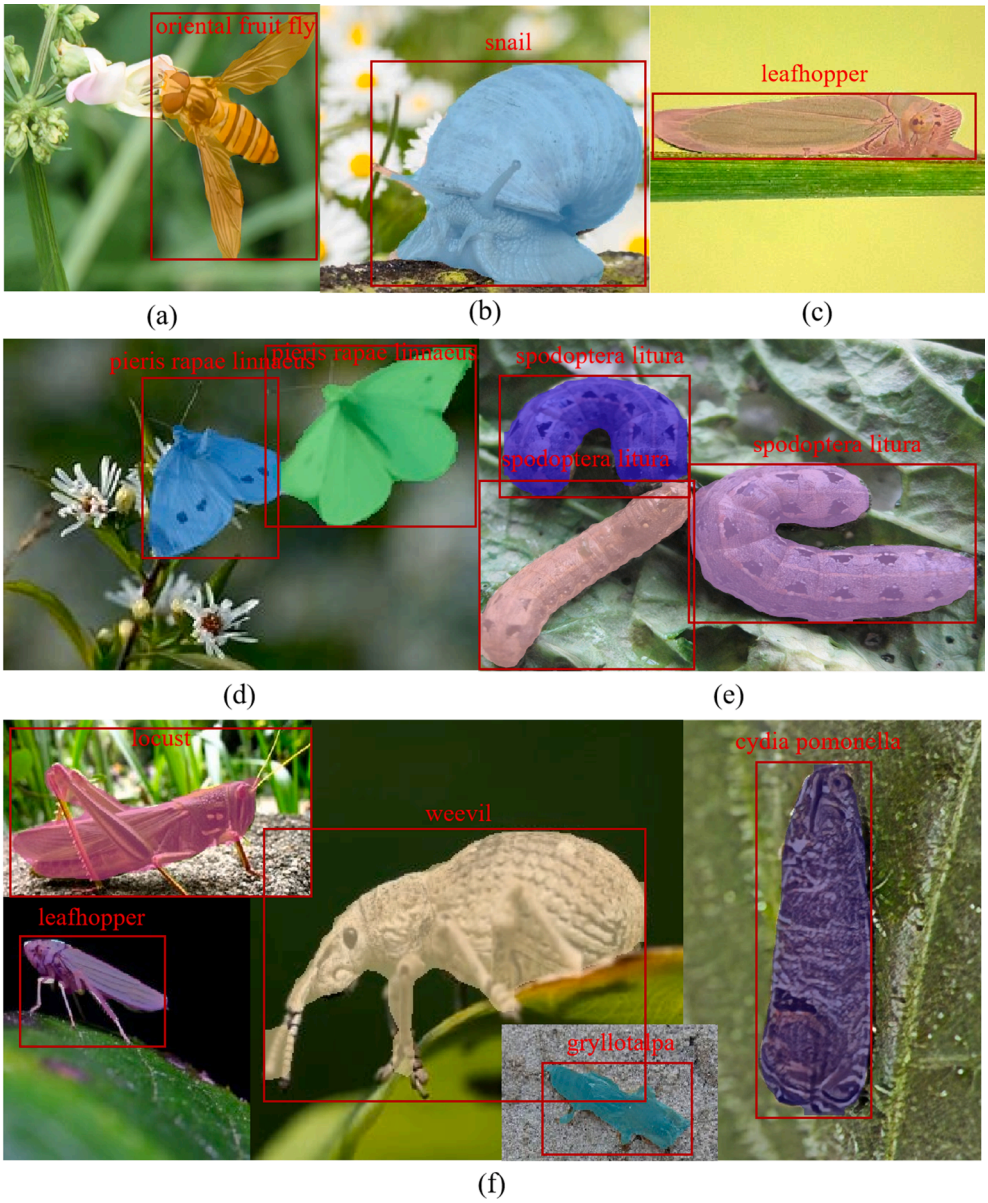

Fig. 10. Detection and segmentation results of the proposed Pest-D2Det model. **Note.** There are three scenarios, which include the images with single pest (a, b, c), the images with multiple pests that belong to the same categories (d, e), and the image with multiple types of pests that belong to different categories (f).

different experimental groups. Table 1 demonstrates the effectiveness of the modified SMSR models on pest detection and segmentation, and the results are assessed by the AP metric. Compared to the fourth experiment without any SR processing, the second experiment that applied SR for training and testing images achieved higher AP values for each class. It suggests that the SR process is necessary for the pest localization framework to enhance the overall performance. Moreover, the third experiment outperformed the fifth experiment for both detection and segmentation tasks. That is probably because the SR operation has a greater impact on the testing process than it does on the training process. Furthermore, the noteworthy gap between the first and second experiments illustrates the improvement of the SMSR model.

5.2. Pest detection and segmentation

In this section, an experiment was conducted to prove that the proposed attention module is of great significance to the feature extraction process in the detection and segmentation tasks. Fig. 8 displays the visualization results of different methods. The first line is the input image, and the second line is the result of the Res2Net101 network without any attention module. The other lines show the results of the networks that integrate different attention modules (CBAM, ECA, and ECSA) with baseline (Res2Net101). Compared with the result of the network without any attention module, the network with an attention module covers more information in the target area. Besides, the ECSA-integrated network (Res2Net101 + ECSA) paid more attention to the pest details like the foot part. That is because the ECSA module learned better and extracted more features from the pest region than the other experimental attention modules.

The modified D2Det was presented by replacing the original base model (ResNet101) in the backbone of the D2Det, and then an effective Pest-D2Det model was proposed by adding a new attention module on the new backbone. In order to assess the Pest-D2Det comprehensively, the model's performances were evaluated in terms of the loss and mAP. Fig. 9 plots the curves of different models, where the blue, orange, and green lines refer to the performances of the original D2Det, the modified D2Det, and the proposed Pest-D2Det models, respectively. Fig. 9 (a) and (b) shows that the Pest-D2Det model obtained the highest detection

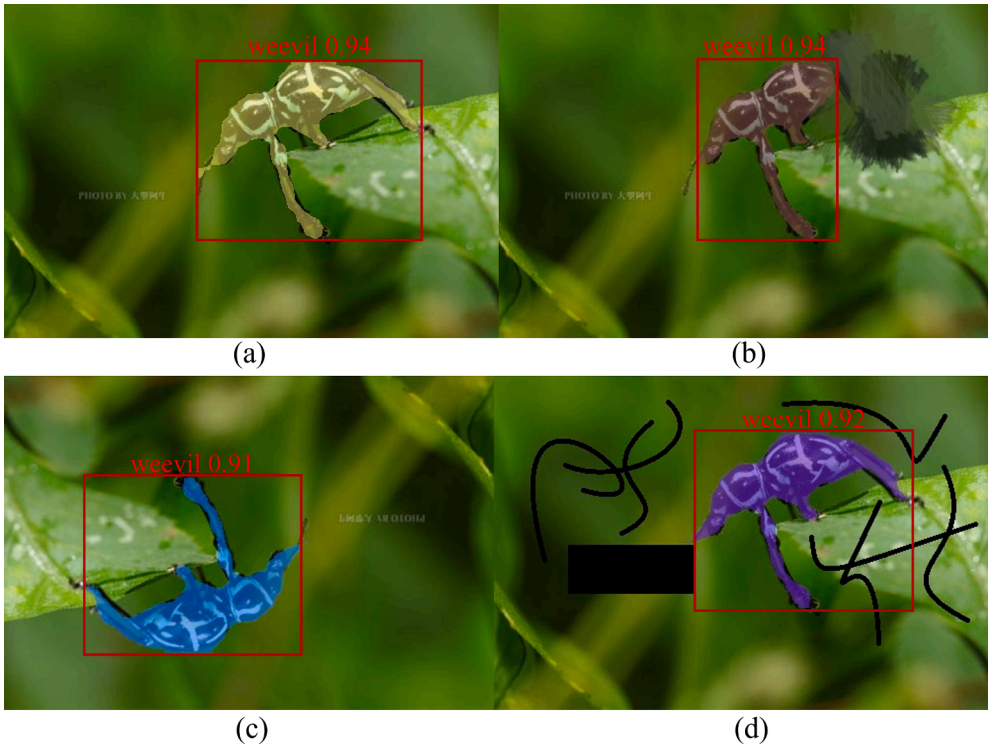

Fig. 11. Detection and segmentation results of the proposed Pest-D2Det model under various challenging conditions. (a) original image, (b) image with mosaic, (c) rotated image, (d) image with man-made noises.

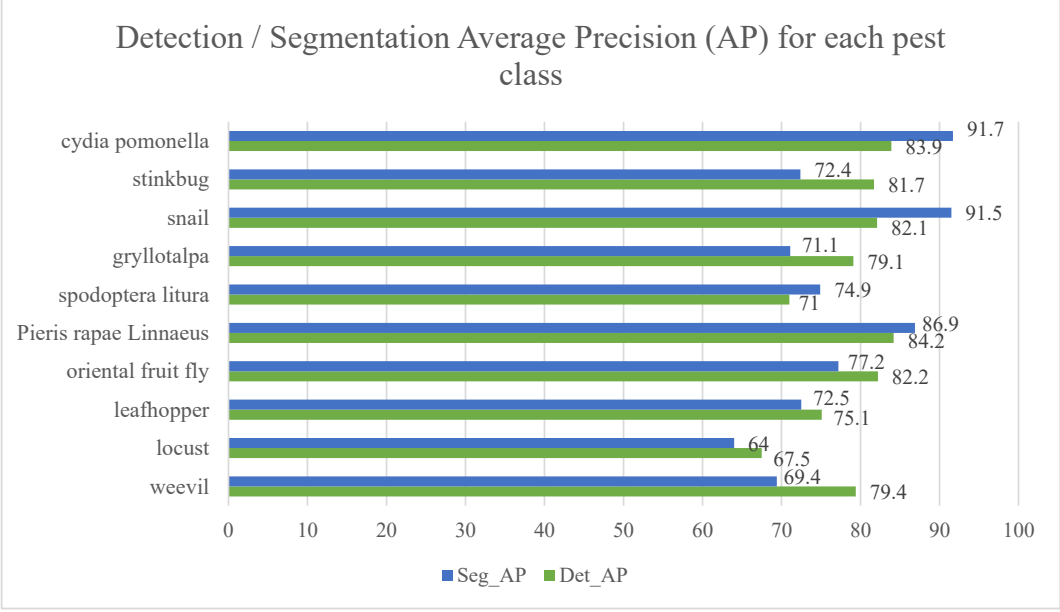

Fig. 12. Detection / segmentation average precision (AP) for each pest class. 'Det_AP' and 'Seg_AP' represent the AP values for pest detection and segmentation, respectively.

mAP of 73.3% and segmentation mAP of 72.6% on the validating set, and the modified D2Det outperformed the original D2Det model. That suggests the modification on the backbone and the proposed attention module can enhance the performance of the original D2Det model. In addition, the loss function curves were used to reflect the training of three experimental models. The loss curve of the original D2Det model shows a significant fluctuation in the first 5000 iterations. In contrast, the loss curve of the Pest-D2Det model has the advantages of stability and fast convergence.

Furthermore, a fair comparison with other top-performing methods is necessary to verify the efficiency and precision of the proposed method by using different backbones in the D2Det framework. As shown in Table 2, parameters and computations of the proposed ECSA module only account for 0.62% and 0.78% of the ECSA-integrated network, respectively. Even the ECA-integrated network has fewer computations (FLOPs) than the proposed network, it is inferior to our network by 1.2% and 1.3% in terms of the values of mAP for both the detection task and the segmentation task. Experiments demonstrate the proposed Pest-

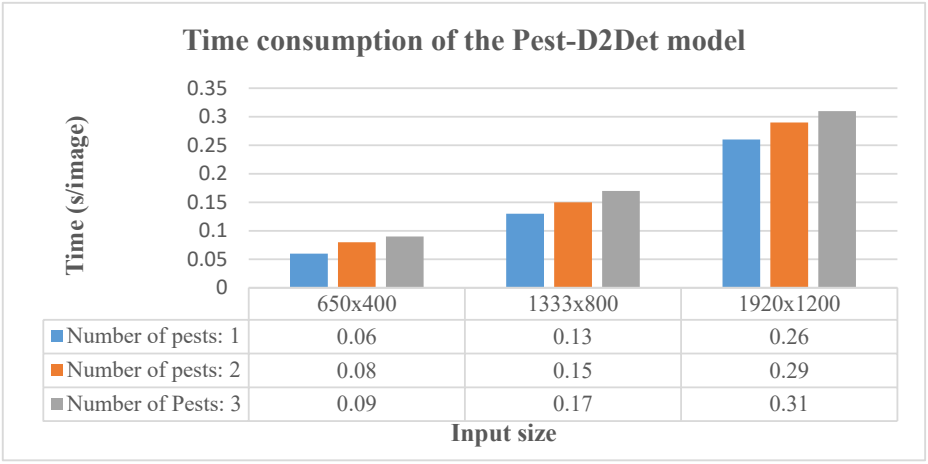

Fig. 13. The time consumption of the Pest-D2Det model on the images with different input sizes and pest numbers.

Table 4
Performance comparison of different instance segmentation models in recent studies based on the same pest dataset.

Model	Input size	Time (s/image)	Det_mAP (%)	Seg_mAP (%)
Mask R-CNN (He et al., 2017)	1333 × 800	**0.114**	66.5	64.7
MS R-CNN (Huang et al., 2019)		0.115	68.2	67.9
HTC (Chen et al., 2019)		0.326	71.8	67.5
D2Det (Cao et al., 2020)		0.149	70.4	69.1
Pest-D2Det		0.157	**78.6**	**77.2**

Note. 'Time' means the average processing time of the whole testing set. 'Det_mAP' and 'Seg_mAP' represent the mAP for pest detection and segmentation, respectively.

D2Det model with the ECSA-Res2Net101 backbone can detect and segment pests with high precision (detection mAP: 73.3%, segmentation mAP: 72.6%) while almost maintaining the original network complexity.

5.3. Parameter analysis

Since fine-tuning approach strengthens the network to learn the features of the targets (Hu et al., 2018), some parameters are adjusted to improve the mAP of the proposed model. Table 3 presents different parameter settings (loss functions of different tasks, optimizers, and initial learning rates) of the Pest-D2Det network and the corresponding performances for both detection and segmentation tasks. A total of four groups of control experiments were conducted to demonstrate the effect of each parameter. According to the experiment 1 and 2, the Pest-D2Det model using the focal loss and CIoU loss achieved better performance. In particular, the segmentation mAP of the 2nd model is 2.2 higher than that of the 1st model. That proved the loss functions used in the 2nd model play a critical role in detecting and segmenting pests. In addition, the detection and segmentation mAP values of the 4th model obtained an improvement of 3.2 and 2.4 compared with the performance of the 2nd model. It is confirmed that the initial learning rate and the optimizer have a clear influence on detection and segmentation.

5.4. Results and discussion

A comprehensive experiment that involves various scenarios is conducted by visualizing and analyzing the results of pest localization and classification. There are three distinct scenarios, from the simple situation to the complicated situation. As shown in Fig. 10, the result images (a, b, c) in the first line contain only one pest, and the images (d, e) in the second line have multiple pests that belong to one class. In the third line, multiple images with a single pest were cropped and stitched into one image, and then the image with various types of pests was input into the proposed model to generate the result image (f). According to the visualized results, it is challenging for the proposed network to segment the pest's tiny body parts like legs and antennae. However, the generated high confidence scores prove that the proposed network can correctly localize and classify the pest images in three different scenarios. As shown in Fig. 10 (c), all body parts of the pest are precisely segmented along its edges. Even if the instances in Fig. 10 (e) are densely distributed and vary in size, they are still correctly segmented with different colors and predicted with the same label. Besides, the image with 5 categories of pests is given different colors and labels in Fig. 10 (f).

Moreover, the robustness of the proposed model is demonstrated by making difficulties on the input images. As shown in Fig. 11 (a), the original input image containing a weevil pest is precisely localized and identified by our method. In Fig. 11 (b), some parts of the pest body are blocked by additional mosaic, and the remaining sections are still localized correctly. Fig. 11 (c) illustrates the proposed method is robust against the image rotation. Finally, in Fig. 11 (d), noises are added manually around the pest, and the noisy image is predicted accurately. It is unambiguous that our network can predict the pest images under various challenging conditions.

On the one hand, the performance of the proposed method is assessed by calculating the detection and segmentation precision of each pest class. As shown in Fig. 12, the proposed model achieved the highest detection AP of 84.2% on the Pieris rapae Linnaeus class, while it obtained the highest segmentation AP of 91.7% on the cydia pomonella class. In contrast, the average precision for the locust class is the lowest among the ten species of pests for both detection and segmentation tasks. The possible reason is that this pest class has many unnoticeable tiny body parts. Besides, the color of this kind of pest is usually similar to the natural background.

On the other hand, the processing time of the proposed model is tested on the images with different input sizes and pest numbers. As shown in Fig. 13, both factors (input size and number of pests) are proportional to the time consumption, whereas the former has a more significant impact on time consumption. As for a single pest image, the processing time of a 1920x1200 image is 0.26 s, which is about four times that of a 650x400 image. That is because the model needs more convolution computation in a large image than in a small image.

5.5. Comparison with other work

In this section, an experiment is constructed to further demonstrate the effectiveness of the proposed model by comparing it with the state-of-the-art. Table 4 shows the compared results of different instance segmentation models in recent research on our collected dataset. Even though the Mask R-CNN model processes per image with the shortest time, its results for detection and segmentation accuracy are unsatisfactory. Compared with the Mask R-CNN, our model obtained a considerable improvement of the detection mAP at 12.1% and the segmentation mAP at 12.5%.

6. Conclusion

This paper proposes an automatic pest localization framework that can efficiently distinguish pest regions from the complicated background. Firstly, a practical and lightweight attention module named ECSA-Net is designed to enhance the learning ability for important features in deep learning-based models. In this study, the proposed ECSA-Net network is integrated into the SR model and the pest monitoring model due to its effectiveness and efficiency. Moreover, an optimized SR method is used for both training and testing images to get better localization and classification results. Furthermore, a dataset including 9,872 images and the corresponding annotation files is fed into a Pest-D2Det network that is proposed by modifying the backbone's structure of the original D2Det model and adjusting parameters. Experiments show that the proposed Pest-D2Det model was robust against various scenarios and achieved state-of-the-art performance in terms of the detection mAP (78.6%) and segmentation mAP (77.2%).

However, the proposed pest monitoring framework cannot be used in a real-time application that requires precise locations of pests in the natural background. Besides, the detection performance of small body parts like legs and antennae is not satisfactory. Therefore, a real-time system will be developed by reducing the computation complexity to perform live videos in the future. In addition, more pest images should be collected and annotated to improve the localization performance of pest details.

CRediT authorship contribution statement

Hanxiang Wang: Conceptualization, Methodology, Data curation, Writing – review & editing. **Yanfen Li:** Conceptualization, Methodology, Data curation, Writing – review & editing. **L. Minh Dang:** Visualization, Investigation. **Hyeonjoon Moon:** Supervision.

Declaration of Competing Interest

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

Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540), National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339) and by a grant(20212020900150) from "Development and Demonstration of Technology for Customers Bigdata-based Energy Management in the Field of Heat Supply Chain" funded by Ministry of Trade, Industry and Energy of Korean governmentand the China Scholarship Council (CSC) (202108260006).

References

Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study. In: in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 126–135.
Araujo, F.H., et al., 2019. Deep learning for cell image segmentation and ranking. Comput. Med. Imaging Graph. 72, 13–21.
Barbedo, J.G.A., 2020. Detecting and Classifying Pests in Crops Using Proximal Images and Machine Learning: A Review. AI 1 (2), 312–328.
Cao, J., Cholakkal, H., Anwer, R.M., Khan, F.S., Pang, Y., Shao, L., 2020. D2det: Towards high quality object detection and instance segmentation. In: in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11485–11494.
Chen, K., et al., 2019. Hybrid task cascade for instance segmentation, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974-4983.
Chen, J.-W., Lin, W.-J., Cheng, H.-J., Hung, C.-L., Lin, C.-Y., Chen, S.-P., 2021. A smartphone-based application for scale pest detection using multiple-object detection methods. Electronics 10 (4), 372.
Dang, L. Minh, et al., 2022. DefectTR: End-to-end defect detection for sewage networks using a transformer. Construction and Building Materials.
Deng, L., Wang, Y., Han, Z., Yu, R., 2018. Research on insect pest image detection and recognition based on bio-inspired methods. Biosyst. Eng. 169, 139–148.
DeVries, T., Taylor, G.W., 2017. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.
Ebrahimi, M.A., Khoshtaghaza, M.H., Minaei, S., Jamshidi, B., 2017. Vision-based pest detection based on SVM classification method. Comput. Electron. Agric. 137, 52–58.
Espinoza, K., Valera, D.L., Torres, J.A., López, A., Molina-Aiz, F.D., 2016. Combination of image processing and artificial neural networks as a novel approach for the identification of Bemisia tabaci and Frankliniella occidentalis on sticky traps in greenhouse agriculture. Comput. Electron. Agric. 127, 495–505.
Gao, S., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., Torr, P.H., 2019. Res2net: A new multi-scale backbone architecture. IEEE Trans. Pattern Anal. Mach. Intell.
He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.
Hong, S.-J., et al., 2020. Moth detection from pheromone trap images using deep learning object detectors. Agriculture 10 (5), 170.
Hu, Y.-T., Huang, J.-B., Schwing, A.G., 2018. Videomatch: Matching based video object segmentation. In: in *Proceedings of the European conference on computer vision (ECCV)*, pp. 54–70.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.
Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X., 2019. Mask scoring r-cnn, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6409-6418.
Jung, A., 2019. Imgaug documentation. Readthedocs. io, Jun, vol. 25.
Kim, J., Lee, J.K., Lee, K.M., 2016. Accurate image super-resolution using very deep convolutional networks. In: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654.
Li, R., et al., 2019. An effective data augmentation strategy for CNN-based pest localization and recognition in the field. IEEE Access 7, 160274–160283.
Li, Yanfen, et al., 2022. A robust instance segmentation framework for underground sewer defect detection. Measurement.
Li, Y., Wang, H., Dang, L.M., Sadeghi-Niaraki, A., Moon, H., 2020. Crop pest recognition in natural scenes using convolutional neural networks. Comput. Electron. Agric. 169, 105174.
Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125.
Liu, L., et al., 2019. PestNet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. IEEE Access 7, 45301–45312.
Maharlooei, M., Sivarajan, S., Bajwa, S.G., Harmon, J.P., Nowatzki, J., 2017. Detection of soybean aphids in a greenhouse using an image processing technique. Comput. Electron. Agric. 132, 63–70.
Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2016. Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440.
Sara, U., Akter, M., Uddin, M.S., 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. J. Comput. Commun. 7 (3), 8–18.
Wang, L., et al., 2021. Exploring Sparsity in Image Super-Resolution for Efficient Inference. In: in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4917–4926.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. ECA-Net: efficient channel attention for deep convolutional neural networks, 2020 IEEE. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
Wang, Z., Wang, K., Liu, Z., Wang, X., Pan, S., 2018. A cognitive vision method for insect pest image segmentation. IFAC-PapersOnLine 51 (17), 85–89.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19.
Yue, Y.i., Cheng, X.i., Zhang, D.i., Wu, Y., Zhao, Y., Chen, Y., Fan, G., Zhang, Y., 2018. Deep recursive super resolution network with Laplacian Pyramid for better agricultural pest surveillance and detection. Comput. Electron. Agric. 150, 26–32.
Zhao, Y., Chen, J., Xu, X., Lei, J., Zhou, W., 2021. SEV-Net: Residual network embedded with attention mechanism for plant disease severity detection. Concurrency and Computation: Practice and Experience 33 (10). https://doi.org/10.1002/cpe.v33.1010.1002/cpe.6161.