European Journal of Agronomy 160 (2024) 127331

Contents lists available at ScienceDirect

European Journal of Agronomy

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

An efficient zero-labeling segmentation approach for pest monitoring on smartphone-based images

L. Minh Dang [a], Sufyan Danish [b], Asma Khan [b], Nur Alam [b], Muhammad Fayaz [b], Dinh Khuong Nguyen [c], Hyoung-Kyu Song [a], Hyeonjoon Moon [b,*]

[a] Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, South Korea
[b] Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea
[c] Department of Mathematics, Faculty of Fundamental Sciences, Van Lang University, Ho Chi Minh City, Vietnam

ARTICLE INFO

Keywords:
Deep learning
Zero-shot
Pest monitoring
Pest recognition

ABSTRACT

Timely and precise farm inspection, which involves the identification and recognition of harmful insects and diseases, is crucial for safeguarding crop production. Traditional vision-based pest recognition methods typically require extensive annotated data for each pest species and a lengthy training process. This approach is time-consuming, labor-intensive, and prone to human error. Zero-shot learning offers a potential solution by enabling pest segmentation and control without requiring explicit training data. This study supports farmers in automatically identifying ten common pests and their precise locations in real-world outdoor environments. The zero-shot pest segmentation is based on a hybrid approach combining Explainable Contrastive Language-Image Pre-training (ECLIP) and Segment-Anything (SAM). Moreover, an optimized super-resolution model and various data augmentation methods are implemented to improve the quality of the dataset. Lastly, a mask post-processing step is applied to remove highly overlapping segmented masks and noise blobs caused by the complex background. The mean Intersection over Union (mIoU) of 66.5 % on the validation set demonstrates the potential of zero-shot methods for automated pest segmentation during farm inspections. This research lays the foundation for accurate pest monitoring systems capable of adapting to new pests, ultimately improving agricultural productivity.

1. Introduction

Crop cultivation faces significant challenges as various pest species inflict damage, which causes declines in both crop yield and quality (Liu and Wang, 2021; Wang et al., 2024). Prompt pest control is crucial in agriculture because it is the most effective way to monitor crop health Jalayer et al. (2023) and optimize yields (Wang et al., 2020; Du et al., 2024). Traditionally, farmers must walk through their fields and search for signs of pest damage such as holes in leaves, discoloration, or wilting Preti et al. (2021). While manual inspection can be effective for detecting large pests, identifying and localizing small or camouflaged pests can be challenging, especially in large fields or complex environments (Sciarretta and Calabrese, 2019; Rustia et al., 2020).

The pressing need for efficient field monitoring in modern agriculture has driven the development of vision-based pest management systems utilizing images captured by cameras, drones, and smartphones Wang et al. (2023). These systems often encounter challenges due to variable image quality, including varying resolutions and lighting conditions, prevalent in real-world agricultural settings. Therefore, a super-resolution (SR) model and various data augmentation techniques are employed to enhance image quality for pest localization in this study Chen et al. (2022). This improvement facilitates the robust recognition of even small or partially occluded pests.

Although current machine learning (ML) and deep learning (DL) techniques have been proven effective for pest management, their performance relies heavily on high-quality curated datasets and lengthy training schedules (Li et al., 2020; Rustia et al., 2021). This reliance limits the scalability and rapid deployment of such technologies in dynamic agricultural environments. With millions of known pest species and new ones being discovered regularly, it is impractical and expensive to collect training data for every pest species Skendžić et al. (2021). By using pre-trained large models to identify and segment pests directly from images, zero-shot approaches can significantly accelerate the pest detection process. Therefore, a comprehensive review conducted by

* Corresponding author.
E-mail address: hmoon@sejong.ac.kr (H. Moon).

https://doi.org/10.1016/j.eja.2024.127331
Received 30 May 2024; Received in revised form 19 August 2024; Accepted 28 August 2024
Available online 3 September 2024

Pourpanah et al. (2022) show that zero-shot learning presents a promising alternative for enhancing overall pest management practices in agriculture, as it does not require explicit training on annotated datasets. In this study, we introduce a novel zero-shot DL-based pest segmentation framework to address the challenges of limited labeled data and complex agricultural environments. The key contributions of our work are as follows:

- An optimized SR algorithm to improve image quality for accurate pest recognition.
- A zero-shot framework combining Segment-Anything (SAM) and Explainable Contrastive Language-Image Pre-training (ECLIP) for effective pest segmentation without requiring labeled data.
- A refined post-processing module to enhance segmentation accuracy by merging overlapping masks and removing noise.

The rest of the paper is organized as follows. Section 2 shows related work, providing a comprehensive overview of previous research in the field of pest segmentation. Section 3 introduces the pest segmentation dataset used in this study, highlighting its characteristics and relevance to the task. Section 4 meticulously outlines the proposed zero-shot pest segmentation pipeline, providing a detailed explanation of its components and their interactions. Section 5 presents a rigorous evaluation of the proposed approach to testing data, examining its performance metrics and discussing the obtained results. Section 6 discusses the zero-shot segmentation framework, exploring its strengths, limitations, and potential applications. Finally, Section 7 concludes the paper with a summary of the key findings and remarks on future directions for research in this domain.

2. Related work

Traditional pest monitoring studies often relied on well-established machine learning (ML) models such as Support Vector Machines (SVMs) (e.g., (Ebrahimi et al., 2017; Ashok et al., 2019)), Decision Trees (e.g., Chopda et al. 2018), and Random Forests (e.g., (Kasinathan and Uyyala, 2021; Zhu et al., 2017)). Although these models achieved good performance in certain scenarios, they suffer from limitations including the requirement for extensive labeled data, sensitivity to environmental changes, and limited ability to capture complex pest-related patterns Domingues et al. (2022). Deep learning (DL) techniques, particularly Convolutional Neural Networks (CNNs) and Transformers, have emerged as promising alternatives. These models excel at processing complex visual data, often requiring large-scale labeled data, and demonstrating superior performance and robustness against real-world field conditions Minh et al. (2022).

Li et al. (2020) demonstrated the effectiveness of a fine-tuned GoogLeNet model in accurately classifying ten pest species with an accuracy of 98 %. Liu et al. (2019) further expanded DL capabilities by introducing PestNet, a novel approach for multi-class pest detection and classification. PestNet consists of three key components: a Channel-Spatial Attention (CSA) module to enhance feature extraction within the CNN backbone, a Region Proposal Network (RPN) to identify potential pest locations, and a Position-Sensitive Score Map (PSSM) to replace traditional fully connected layers for classification and bounding box regression. The experimental results highlight PestNet's superior performance in multi-class pest detection, achieving a mean Average Precision (mAP) of 75.46 %. It also achieves good performance on pest classification with an overall accuracy of 92.3 %.

Deep learning-based pest monitoring frameworks rely on high-quality and sufficient data for accurate detection and recognition. To address challenges from real-world image variations, such as scale, posture, noise, and blur, researchers have explored image processing techniques Dang et al. (2024). Data augmentation Shorten and Khoshgoftaar (2019) has been employed to mitigate issues related to scale and posture, while super-resolution (SR) methods Wang et al. (2021) have been utilized to enhance image clarity and improve pest detection in low-resolution images.

Due to the complicated nature of pest identification, zero-shot learning has been increasingly applied to perform automated pest monitoring. Zero-shot learning is a recent ML trend that allows models to generalize to new data distributions and tasks without explicit training examples. The Segment Anything Model (SAM) Kirillov et al. (2023) introduced by Meta AI is a pre-trained large model that can segment objects without prior training on specific data distributions. This flexibility allows SAM to handle a wide variety of segmentation tasks, even when dealing with novel tasks. Contrastive Language-Image Pre-training (CLIP) Radford et al. (2021) is another promising approach for zero-shot learning. CLIP's ability to associate text and image representations enables it to recognize and understand novel objects without extensive labeled data. SAM and CLIP models align well with the challenges of pest identification, where data collection for numerous species is time-consuming and costly. A recent study by Zhong et al. (2020) introduced a conditional adversarial autoencoder (CAAE)-based approach for zero- and few-shot learning for disease recognition. A zero-shot disease classifier, trained on a small number of labeled samples from seen classes, is used to classify unseen classes by transferring the knowledge learned from the CAAE model. In contrast, a few-shot disease classifier, trained on a small number of labeled samples from both predefined and unseen classes, utilizes the knowledge learned from the CAAE model and the few labeled samples to classify unseen classes. The experimental results indicate that the proposed method outperformed state-of-the-art models in zero-shot disease classification with a mean accuracy of 53.4 %.

Traditional image processing and ML methods were favored for their ease of implementation and low computational complexity. However, these methods relied heavily on handcrafted features. On the other hand, DL models offer superior accuracy but demand substantial data and computational resources. This study introduces a novel zero-shot pest segmentation framework to overcome the limitations of previous approaches and enable efficient pest recognition in natural environments. Moreover, data augmentation techniques and a sparse convolution-based super-resolution model are employed to improve the dataset quality.

3. Pest monitoring dataset

3.1. Data collection

A dataset comprising 2000 images was curated from the crop pest database of Li et al. (2020) through manual selection. Image acquisition involved both online sourcing from popular search engines (Google, Baidu, Yahoo, and Bing) and outdoor mobile phone photography using an Apple 7 Plus. The dataset contains ten distinct pest species: codling moth (*Cydia pomonella*), mole cricket (*Gryllotalpa gryllotalpa*), leafhoppers (*Cicadellidae*), locust (*Locusta migratoria*), fruit fly (*Bactrocera dorsalis*), cabbage white (*Pieris rapae*), snail (*Helix aspersa*), cotton leafworm (*Spodoptera littoralis*), stinkbug (*Halyomorpha halys*), and weevil (*Curculionidae*). Images exhibit diverse resolutions ranging from 224 × 90–4000 × 2337 pixels, and most were captured in real-world scenarios. The dataset was randomly divided into training and testing sets with a ratio of 8:2, resulting in 1600 training images and 400 testing images. To obtain the segmentation mask, all pest images were labeled with the publicly available annotation tool LabelMe Russell et al. (2008). The generated segmentation masks are in JSON format, with annotated pixels indicating the pest's location. Sample images for each pest category from the dataset are presented in Figure 1.

3.2. Data augmentation

The experimental section evaluates the zero-shot approach against state-of-the-art supervised methods. Unlike zero-shot methods,

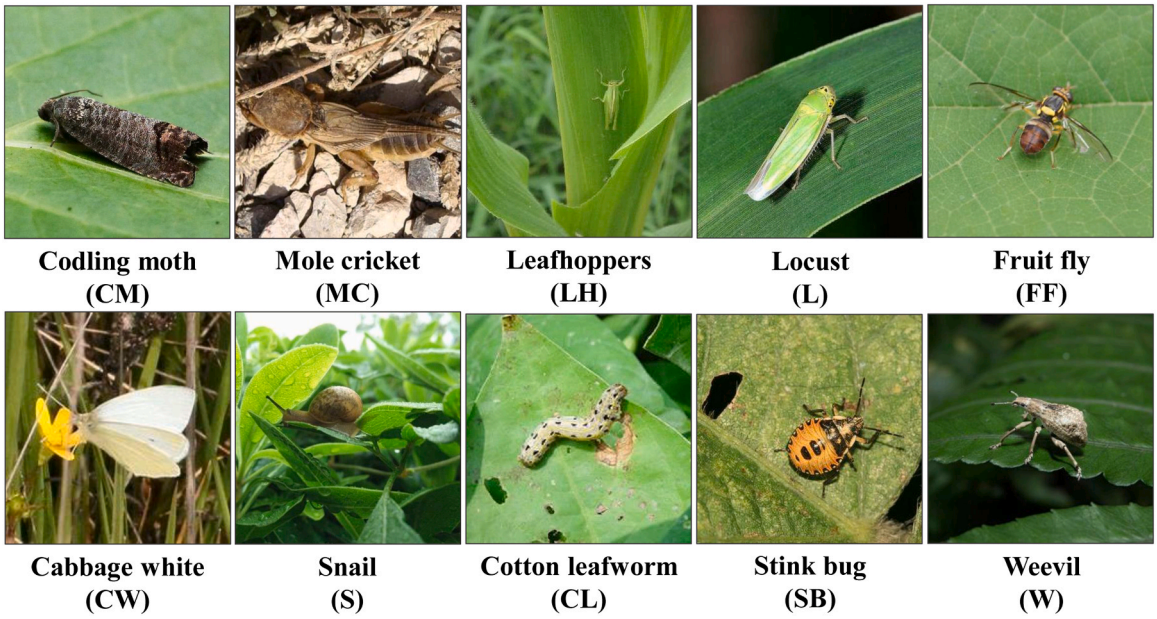

Fig. 1. Representative images of the ten pest species included in the dataset.

supervised approaches require substantial annotated training data. To enhance data diversity, we employed data augmentation techniques including gamma contrast adjustment, random flips, rotations (− 20 to 20 degrees), Gaussian blur, brightness modifications, channel shuffling, and cutout. These augmentations simulated real-world image variations such as lighting conditions, object orientations, and occlusions. For instance, channel shuffling introduced color variations while cutout operations mimicked obscured pest regions.

To increase the training dataset size, the mentioned data augmentation techniques were applied to the original training set of 1600 images. After the augmentation process, the training set was expanded to 9840 images—a sixfold increase. Figure 2 illustrates examples of the applied data augmentation methods.

4. Methodology

4.1. Image pre-processing

Previous research highlights the challenges posed by poor lighting, low resolution, and low contrast in pest datasets due to real-world imaging conditions, which can significantly degrade the performance of pest identification models (Liu and Wang, 2021; Ashok et al., 2019). Therefore, SR methods are typically integrated into the preprocessing step to improve image quality. This study employs Sparse Mask Super-resolution (SMSR), a state-of-the-art SR model, proposed by Wang et al. (2021) to upscale pest images.

The SMSR model accelerates image super-resolution by identifying and skipping redundant computations in image SR networks. The SMSR model consists of two main components: a spatial mask learning module and a channel mask learning module. The spatial mask learning module is implemented by training a CNN to predict a binary mask for the input image, where 1 indicates a redundant region and 0 indicates a non-redundant region. Similarly, the channel mask module, also based on a CNN, identifies redundant feature map channels, where 1 indicates a redundant channel and 0 indicates a non-redundant channel. Once the spatial and channel masks have been learned, they are used to skip redundant computations in the SR network. For example, if a region in the input image is identified as redundant, the SR network does not need to compute any features for that region. Similarly, if a channel in the feature maps is identified as redundant, the SR network does not need to compute any outputs for that channel.

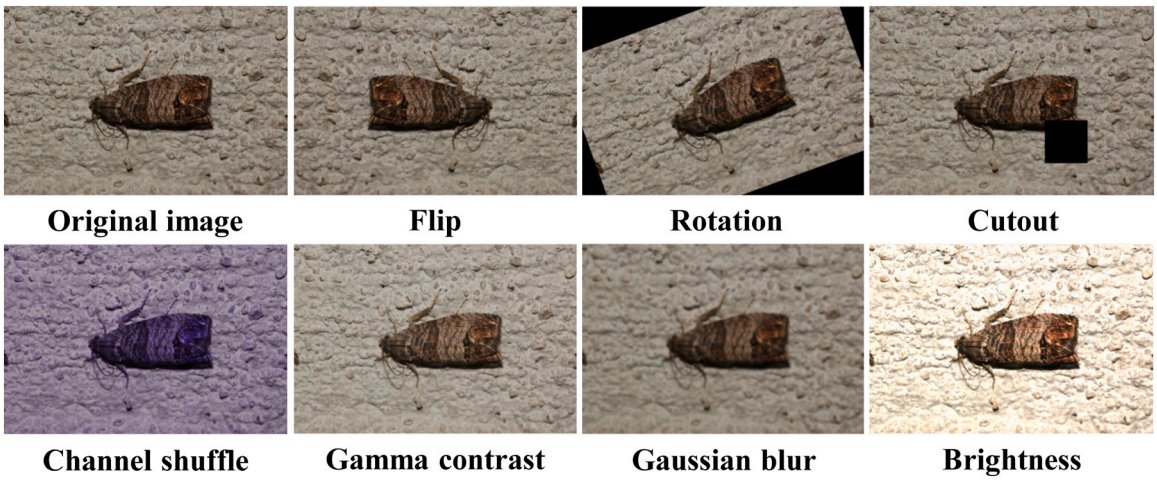

Fig. 2. Augmented image samples generated using seven different augmentation techniques.

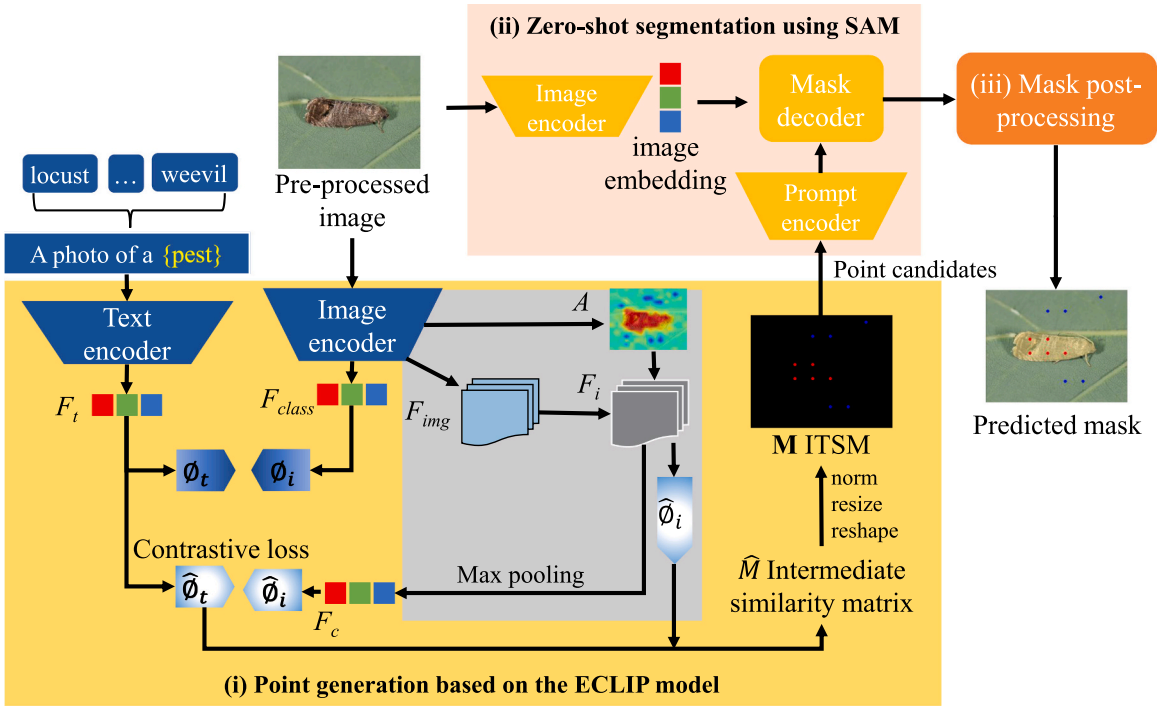

Fig. 3. Visualization of the text-to-points candidate generation process using ECLIP to guide SAM in generating the mask for an input pest image. **Note:** Class features (F_{class}), image features (F_{img}), text features (F_t), pooled features (F_c), masked features (F_i), and an expanded mean attention map (A). Moreover, the framework incorporates dual projections, ϕ_i and $\hat{\phi}_i$, along with their corresponding text projections, ϕ_t and $\hat{\phi}_t$, which are utilized for calculating the contrastive losses.

4.2. Zero-shot pest segmentation framework

Figure 3 illustrates the three core components of the proposed zero-shot pest segmentation framework.

- Point candidate generation: The Explainable contrastive language-image pre-training (ECLIP) model generates point candidates for pests based on input images and corresponding pest text descriptions. These point candidates are later used to guide the SAM segmentation model towards relevant pest regions in the image.
- Zero-shot pest segmentation: Using the point candidates generated by ECLIP, SAM effectively identifies the target pest in the image and produces initial masks outlining the target pest's boundaries.
- Mask post-processing: The initial segmentation masks may exhibit duplicates and noise blobs due to the complex background. Therefore, to refine the masks and obtain precise and detailed pest segmentation results, a post-processing module is employed to eliminate duplicate masks and reduce noise.

4.2.1. Point candidate generation

Contrastive Language-Image Pre-training (CLIP), introduced by Radford et al. (2021), has significantly advanced multimodal representation learning by jointly embedding images and text descriptions into a shared latent space. This shared space ensures that similar images and their associated text descriptions are mapped to nearby points, which fosters strong correlations between visual and linguistic information. CLIP demonstrates exceptional performance across diverse downstream vision tasks, including object segmentation Kirillov et al. (2023), image retrieval Saito et al. (2023), and zero-shot classification Wei et al. (2023). CLIP's ability to recognize novel object categories without explicit training highlights its powerful generalization capabilities. However, CLIP operates as a black box, and explaining its internal mechanism remains an open research question.

Li et al. [2022] introduced ECLIP, an explainable variant of CLIP that enhances model interpretability through feature map visualization. By incorporating an Image-Text Similarity Map (ITSM), ECLIP quantifies the similarity between image features and corresponding text descriptions. ITSM provides valuable insights into the model's decision-making process by identifying image regions most relevant to the given text. To further enhance interpretability, ECLIP replaces the standard global pooling layer with a masked max pooling layer. This modification focuses attention on image regions deemed relevant to the provided text description by the ITSM. By concentrating on these relevant regions, the masked max pooling layer effectively reduces noise and improves model explainability. The combined use of ITSM and masked max pooling enables ECLIP to generate visually interpretable explanations for its predictions.

Given an input pest image x and its corresponding text description y, ECLIP employs supervised image and text encoders, f_i and f_t, respectively, along with a corresponding pair of linear projections head ϕ_i and ϕ_t. The image encoder processes the input image to generate image features X, containing a class token $X_c \in \mathbb{R}^{1\times C}$ for classification and image tokens $X_i \in \mathbb{R}^{N_i\times C}$ representing the raw feature map. Concurrently, the text encoder processes the text description to produce text features $Y \in \mathbb{R}^{N_t\times C}$, as shown in Equation (1).

$$X = f_i(x), Y = f_t(y) \tag{1}$$

where C denotes embedding size, the class token is represented by a single vector, and N_i and N_t represents the number of image and text tokens, respectively.

As described in Equation (2), an intermediate similarity matrix $\hat{M} \in \mathbb{R}^{N_i\times N_t}$ is computed by calculating the inner product between the image token X_i (excluding the class token X_c) and the text features Y.

$$\hat{M} = \left(\frac{X_i\cdot\phi_i}{\|X_i\cdot\phi_i\|_2}\right)\cdot\left(\frac{Y\cdot\phi_i}{\|Y\cdot\phi_i\|_2}\right)^{\top} \tag{2}$$

To enhance visual interpretability, the ITSM feature map, $M \in \mathbb{R}^{H\times W\times N_t}$, undergoes a transformation process involving resizing and min-max normalization. Initially, the feature map is resized to match the input image dimensions using bicubic interpolation. This ensures that the spatial arrangement of the similarity scores corresponds accurately to the layout of the image. After that, min-max normalization is applied

to the feature maps to ensure that the input is more interpretable and suitable for visualization. The entire transformation can be expressed as:

$$\mathbf{M} = \text{Norm}(\text{Resize}(\text{Reshape}(\widehat{M}))) \quad (3)$$

Here, W and H denote the width and height of the input image, respectively. The transformed matrix M provides a visually interpretable representation of similarity scores between image regions and the text description. This enhanced visualization facilitates the identification and interpretation of the image regions most relevant to the textual content.

In the context of zero-shot pest segmentation in this study, points with similarity scores exceeding 0.8 are selected as candidate points, and an equal number of lowest-ranked points are selected as background. The candidate points extracted from the ECLIP model are then used to guide SAM in segmenting potential pest regions in the image without the need for an annotated dataset

4.2.2. Zero-shot segmentation

Building upon recent advancements in prompt-based semantic segmentation, this study employs the SAM model introduced by Kirillov et al. (2023) as the zero-shot pest segmentation model. Unlike traditional methods that rely on extensive labeled data, SAM achieves accurate object segmentation through concise textual or point candidates. Trained on the massive SA-1B dataset of image-text pairs, SAM develops a profound understanding of visual-linguistic relationships, enabling it to segment objects, including unseen categories, with remarkable performance and flexibility.

As described in Figure 3(ii), the SAM model comprises three primary components: an image encoder, a prompt encoder, and a mask decoder. Each component is described as follows.

- Image encoder: The image encoder serves as the backbone of the SAM model. It utilizes Vision Transformers (ViTs) Dosovitskiy et al. (2020) to partition the input image into smaller, non-overlapping regions (patches). Feature extraction is then performed to extract crucial features from each individual patch. This strategy allows ViTs to capture both fine-grained details and global context effectively.
- Prompt encoder: The prompt encoder shows remarkable versatility because it can handle both sparse (points, boxes, and texts) and dense (masks) prompts. In the context of pest segmentation, where precise pest locations are unknown, the SAM model utilizes candidate points generated by the ECLIP model to guide the segmentation process. The prompt encoder effectively transforms these points into a latent representation capturing essential pest-related information.
- Mask decoder: The mask decoder integrates encoded image features with the latent representation derived from the prompt encoder. This fusion allows the model to precisely identify pest regions within the image. The mask decoder then generates a segmentation mask that accurately outlines the boundaries of the pests.

4.2.3. Mask post-processing

Masks generated from point candidates often contain duplicates and noise due to complex backgrounds. To address this, a mask post-processing algorithm (Algorithm 1) was implemented to refine segmentation results Nguyen et al. (2023). This process eliminates excessively large or small masks and merges overlapping masks based on predefined intersection over union (IoU) and overlap ratio thresholds. This refinement leads to cleaner and more accurate segmentation outputs.

Input: sam_generated_masks, min_area, max_area, iou_threshold, overlap_threshold.

```
filtered_masks ← []
for each mask in sam_generated_masks do
    mask, mask_area ← find_largest_contour(mask)
    if min_area ≤ mask_area ≤ max_area then
        filtered_masks ← filtered_masks ∪ mask
    end if
end for
processed_masks ← []
while filtered_masks ≠ ∅ do
    pivot_mask ← filtered_masks.pop()
    for each mask in filtered_masks do
        iou, overlap_ratio ← calc_mask_overlap(pivot_mask, mask)
        if (iou > iou_threshold) or (overlap_ratio > overlap_threshold)
    then
            pivot_mask ← pivot_mask ∪ mask
        end if
    end for
    processed_masks ← processed_masks ∪ pivot_mask
end while
```

5. Experimental results

This section presents a comprehensive evaluation of the proposed zero-shot pest segmentation framework using the pest dataset as a benchmark. Section 5.1 outlines the evaluation metrics used to assess model performance, while Section 5.2 describes the hardware and software environment utilized for model development.

5.1. Evaluation metrics

The confusion matrix is a fundamental evaluation metric for semantic segmentation tasks. It provides a detailed breakdown of model performance on different classes by quantifying correctly and incorrectly classified pixels. Therefore, the confusion matrix enables a comprehensive assessment of the model's ability to distinguish between distinct object categories within an image.

- True positive (TP): Pixels correctly predicted to belong to a specific class.
- False positive (FP): Pixels incorrectly predicted to belong to a certain class.
- False negative (FN): Pixels incorrectly predicted to not belong to a specific class.

The confusion matrix provides the foundation for calculating Intersection over Union (IoU), a standard metric for evaluating segmentation performance. IoU quantifies the overlap between predicted and ground truth segmentation masks, with higher values indicating better alignment between the masks. To assess performance across multiple classes, mean IoU (mIoU) is computed by averaging IoU values for each class. The following equations define IoU and mIoU:

$$\begin{aligned} \mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{mIoU} &= \frac{\mathrm{IoU}}{\mathrm{N}} \end{aligned} \tag{4}$$

where N, which represents the total number of pest classes in this study, is set to 10.

5.2. Implementation details

The proposed zero-shot pest segmentation framework was developed using PyTorch,[1] a popular DL library for Python. Training and evaluation were conducted on a Linux machine equipped with two NVIDIA Tesla V100 GPUs, each having 32 GB of memory. The evaluation of the model's performance was conducted exclusively on the original images in the validation set to ensure that the evaluation results accurately reflect the model's performance in real-field scenarios, where data augmentation is not applicable. To establish a benchmark for our zero-shot approach, three supervised segmentation models were implemented. SegFormer and DeepLabv3 were employed using MMSegmentation Contributors (2020), an open-source PyTorch-based toolbox and benchmark for semantic segmentation tasks. Pest-D2Det was implemented based on the implementation details described in the original paper Wang et al. (2022).

5.3. Evaluation of the Zero-Shot Pest Segmentation Framework's Performance

5.3.1. Pre-processing module analysis

The image preprocessing stage involves applying the SMSR super-resolution model to enhance the resolution of the raw pest dataset. The effectiveness of the SMSR algorithm was evaluated by calculating the Peak Signal-to-Noise Ratio (PSNR). Figure 4 illustrates SMSR outputs for stink bug and cotton leafworm images at a 4 × scale factor. By comparing regions of interest (ROIs) from ground truth high-resolution images, low-resolution inputs, and SMSR output images, the model's ability to recover fine details, such as leaf textures and pest shapes, from severely degraded inputs becomes evident. The SMSR model shows superior performance in reconstructing high-quality images from low-resolution counterparts and preserving image clarity and detail, as demonstrated by visual comparisons and PSNR values exceeding 32 in both cases.

To comprehensively assess the impact of pre-processing, Table 1 compares the performance of four segmentation models (zero-shot (ECLIP+SAM), SegFormer, Pest-D2Det, and DeepLabv3) using both raw and pre-processed data based on three segmentation metrics: mIoU, precision, and recall. For the zero-shot approach, the application of pre-processing techniques resulted in notable improvements in all metrics, with mIoU increasing from 62.6 % to 66.5 %, precision from 61.5 % to 68.8 %, and recall from 63.1 % to 67.9 %. These results indicate the importance of image quality enhancement for accurate pest segmentation, particularly in the context of zero-shot learning.

Similarly, supervised models also show performance improvements with pre-processed data. SegFormer achieved a small improvement in mIoU from 73.1 % to 73.8 %. Pest-D2Det demonstrated more significant improvements, with mIoU increasing from 73.8 % to 75.6 %, precision from 74.2 % to 76.4 %, and recall from 74.3 % to 77.1 %. DeepLabv3, while initially lagging in performance with raw data, showed better performance with pre-processed data, with mIoU increasing from 65.4 % to 68.2 %, precision from 68.2 % to 72.1 %, and recall from 66.5 % to 69.3 %. These results collectively suggest the critical role of the SMSR model in improving data quality for the pest segmentation networks. The enhanced image quality likely enabled the model to identify and segment the target pests within the images more accurately. Furthermore, the SMSR model's ability to eliminate noise and artifacts from the test images likely contributed to improved pest segmentation. Finally, the SMSR model enhanced contrast and color in the input images, further assisting the segmentation model in identifying various types of pests.

5.3.2. Detailed evaluation of the zero-shot pest segmentation framework

Table 2 presents the performance of the zero-shot pest segmentation framework for each pest type in the dataset. The model consistently achieves impressive IoU, precision, and recall scores exceeding 60 % across all ten pest categories. These results highlight the framework's robustness in handling real-world challenges, including varying lighting conditions, pest appearances, image quality, and occlusions.

The highest IoU values are obtained for codling moth (70.1 %), cabbage white (69.5 %), and stink bug (67.7 %), while the lowest IoU scores are recorded for leafhoppers (61.3 %) and mole crickets (62.8 %). A potential explanation for the comparatively low segmentation performance of leafhoppers and mole crickets is that they can be difficult to distinguish due to their colors blending with their surroundings. For example, leafhoppers often have green coloration that blends in with the foliage of plants, while mole crickets have brown or gray coloration that makes them difficult to see against the soil. Precision generally outperforms recall across most classes, which indicates the model excels at identifying true positives, but might miss some instances, leading to higher false negatives. Despite these challenges, the proposed zero-shot pest segmentation approach achieves good performance on all ten pest types. Therefore it has a huge potential for practical applications in precision agriculture.

Figure 5 illustrates the proposed model's ability to generate accurate segmentation masks for ten target pest classes. The first column presents the original images for analysis. The second column displays the attention masks generated by ECLIP, which highlight regions within the image that the model identifies as potentially containing pests. The third

[1] https://pytorch.org/

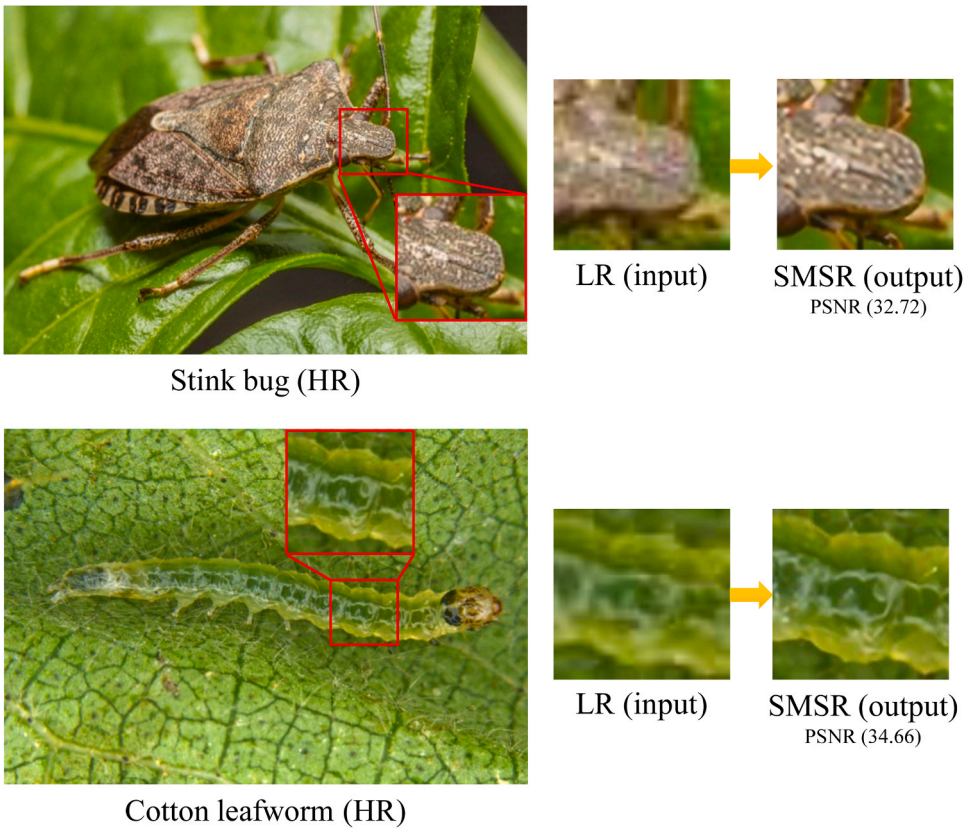

Fig. 4. Visualization of the super-resolution results using the SMSR model with a 4 × scale factor. Note: Comparison of low-resolution (LR) input, SMSR output (SR) output, and ground truth images (HR). **Note:** LR is low-resolution, SR is super-resolution, and PSNR is Peak Signal-to-Noise Ratio.

Table 1
Analysis of the positive impact of pre-processing module on model performance.

Model	Approach	mIoU (%)	Precision (%)	Recall (%)
SegFormer	Raw data	73.1	72.3	73.5
	Pre-processed data	73.8	75.2	75.4
Pest-D2Det	Raw data	73.8	74.2	74.3
	Pre-processed data	75.6	76.4	77.1
DeepLabv3	Raw data	65.4	68.2	66.5
	Pre-processed data	68.2	72.1	69.3
Our (ECLIP+SAM)	Raw data	62.6	61.5	63.1
	Pre-processed data	66.5	68.8	67.9

column overlays the predicted pest masks onto the corresponding original images. These visualizations demonstrate the effectiveness of ECLIP's attention mechanism in identifying potential pest locations within the images, even under challenging conditions with noise and occlusion. This is particularly evident in the cases of locusts and mole crickets, where the attention masks accurately pinpoint the pests despite their seamless visual blend with their surroundings, such as leaves and soil. Similarly, the attention masks successfully recognize tiny pests, such as leafhoppers and weevils. Guided by these attention masks, point candidates are extracted to help SAM generate precise segmentation masks by accurately tracing pest boundaries.

Figure 6 shows the model's robustness in handling three different challenging cases. Subfigure 6(a) reveals the model's ability to recognize and segment small pests, specifically leafhoppers in this case. Despite being tiny, at a long distance from the camera, and camouflaged due to color similarity with the surrounding plant leaves, the model successfully detected and segmented the leafhoppers. Subfigure 6(b) highlights the model's ability to handle occlusion. The model effectively localized the entire mole cricket, even when a portion of its body was hidden underground. Subfigure 6(c) presents a case of multiple cotton worms. Although the model successfully segmented most individuals, it made minor segmentation errors in parts of some worms resembling leaf shadows (red arrows).

5.3.3. *Comparison analysis of the zero-shot segmentation*

To evaluate the performance of the zero-shot pest segmentation framework against recent supervised segmentation models (Deeplabv3 Chen et al. 2017, Pest-D2Det Wang et al. 2022, and SegFormer Xie et al. 2021). All supervised models used the same pre-processed dataset as the proposed zero-shot approach to ensure consistency and fairness across subsequent experiments. Moreover, a pre-trained ResNet-50 model on ImageNet served as the backbone for all supervised models. Model architectures and hyperparameters were adopted from their respective original publications, excluding the proposed zero-shot segmentation model.

Table 3 presents a comprehensive performance comparison across

Table 2
Performance comparison of the zero-shot segmentation approach across ten different pest classes.

	CM	MC	LH	L	FF	CW	S	CL	SB	W
IoU	70.1	62.8	61.3	65.7	68.2	69.5	67.1	66.4	67.7	66.3
Precision	72.6	63.8	63.7	68.5	71.1	71.2	70.4	68.4	69.2	69.9
Recall	71.2	63.9	62.3	67.8	70.5	72.7	69.7	67.1	66.4	67.5

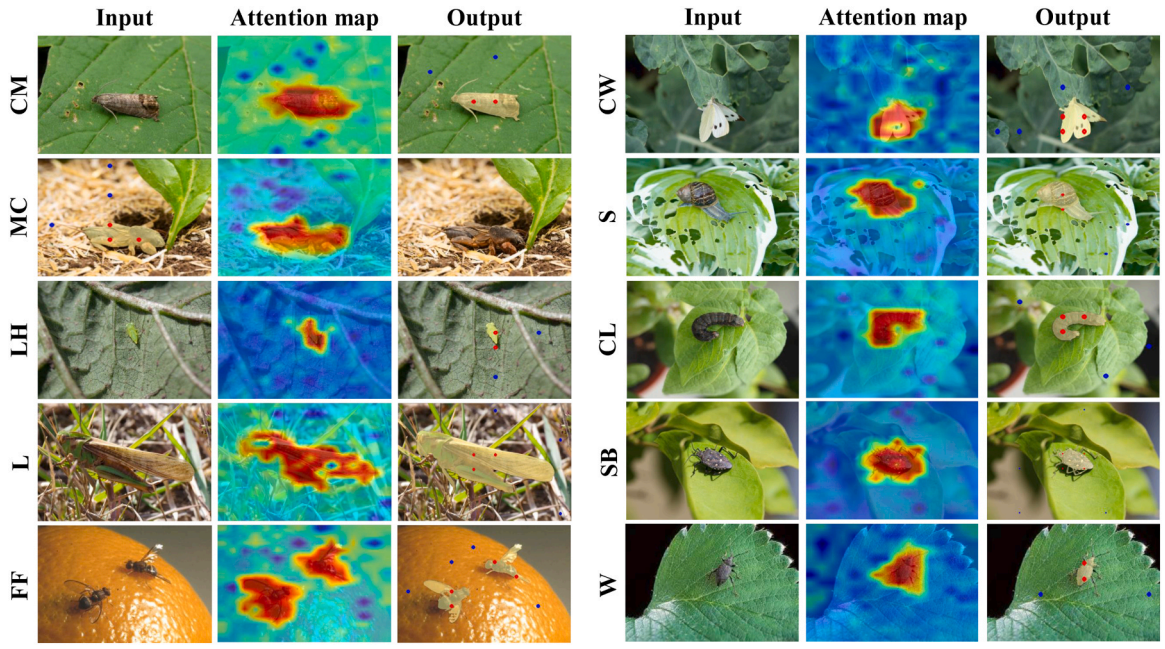

Fig. 5. Attention maps and segmentation results visualization for ten pest types using the proposed zero-shot framework. **Note:** In each image of the third column, red dots indicate regions identified by the model as potential pest locations, while blue dots represent background areas. This dual visualization scheme helps guide the SAM model.

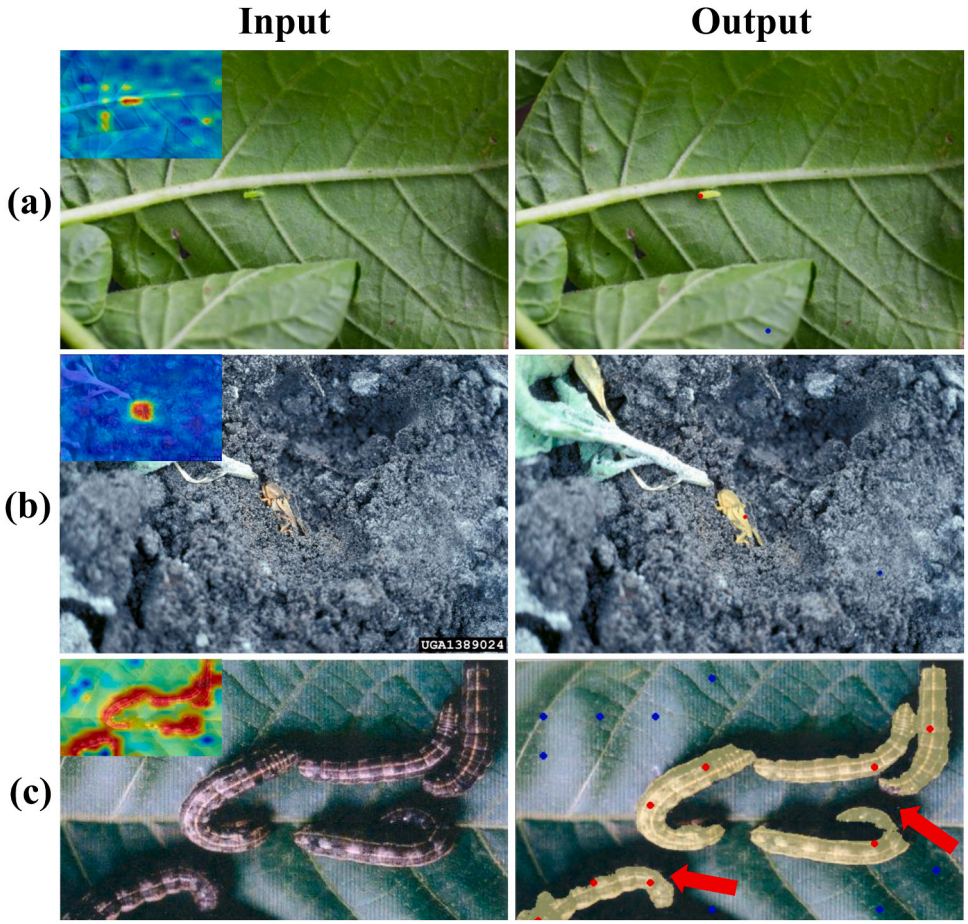

Fig. 6. Zero-shot pest segmentation model performance on challenging scenarios: small object, occlusion, and multiple instances. **Note:** The model's prediction process incorporates two key outputs for each image: attention maps and segmentation results. The red arrows indicate some mis-segmentation areas.

Table 3
Performance comparison of zero-shot and supervised segmentation models on the annotated testing dataset.

Approach	mIoU	Precision	Recall	FPS
DeepLabv3 Chen et al. (2017)	68.2	72.1	69.3	**19**
Pest-D2Det Wang et al. (2022)	**75.6**	**76.4**	**77.1**	14
SegFormer Xie et al. (2021)	73.8	75.2	75.4	11
Ours (ECLIP+SAM)	66.5	68.8	67.9	7

Table 4
Analysis of segmentation performance using five-fold cross-validation on 10 classes of the pest segmentation dataset.

Fold	Zero-shot approach	SegFormer	Pest-D2Det	DeepLabv3
1	67.1	72.1	74.3	68.7
2	64.9	74.4	77.1	68
3	66.8	73.2	74.6	66.5
4	66.3	75.8	75.9	68.8
5	64.7	73.5	74.2	67.4
Mean	65.96	73.8	75.22	67.88

four key metrics: mIoU, precision, recall, and frames per second (FPS). Higher mIoU values denote better segmentation accuracy due to the greater degree of overlap between the predicted and ground truth segmentation masks. Precision and recall, on the other hand, quantify the model's ability to correctly identify true positives (pest pixels) and minimize false positives (background pixels) and false negatives (missed pest pixels), respectively. Finally, FPS serves as a metric for processing speed, with higher values indicating faster model inference speed.

Among the supervised segmentation models, Pest-D2Det achieves the highest mIoU of 75.6 %, followed by a precision of 76.4 % and recall of 77.1 %. Pest-D2Det is a model designed specifically for pest segmentation, which demonstrates superior segmentation accuracy and effectiveness in correctly identifying pest regions. While the model achieves a high level of accuracy, its inference speed of 17 FPS presents a limitation for real-time applications. This suggests its suitability for scenarios where time is not a critical factor, such as offline supervised pest segmentation tasks. SegFormer also performs well, with an mIoU of 73.8 %, precision of 75.2 %, and recall of 75.4 %, although it is notably slower with an FPS of 11. DeepLabv3 shows balanced performance with an mIoU of 68.2 %, precision of 72.1 %, and recall of 69.3 %, but stands out with the highest FPS of 19.

In comparison, the zero-shot approach (ECLIP+SAM) demonstrates lower performance metrics with an mIoU of 66.5 %, precision of 68.8 %, recall of 67.9 %, and a low FPS of 7. These results suggest that while the zero-shot method offers a competitive alternative without the need for extensive annotated training data, it currently lags behind supervised methods in terms of segmentation accuracy and processing speed. The lower FPS indicates slower inference times, which may impact its practicality in real-time applications. However, the zero-shot approach's ability to function without training on specific datasets still highlights its potential for scenarios where annotated data is scarce or unavailable. This comparison emphasizes the trade-offs between supervised and zero-shot methods and the need for further optimization in zero-shot techniques to bridge the performance gap.

Table 4 describes a five-fold cross-validation experiment conducted on the pest segmentation dataset to evaluate the robustness and generalizability of the proposed zero-shot model and compare it to supervised counterparts, including SegFormer, Pest-D2Det, and DeepLabv3. Each fold involved training on four subsets and validating the remaining subset. In contrast, the zero-shot model was evaluated solely on the validation subset due to its training-free nature. The cross-validation strategy mitigates the potential bias of a single data partition and

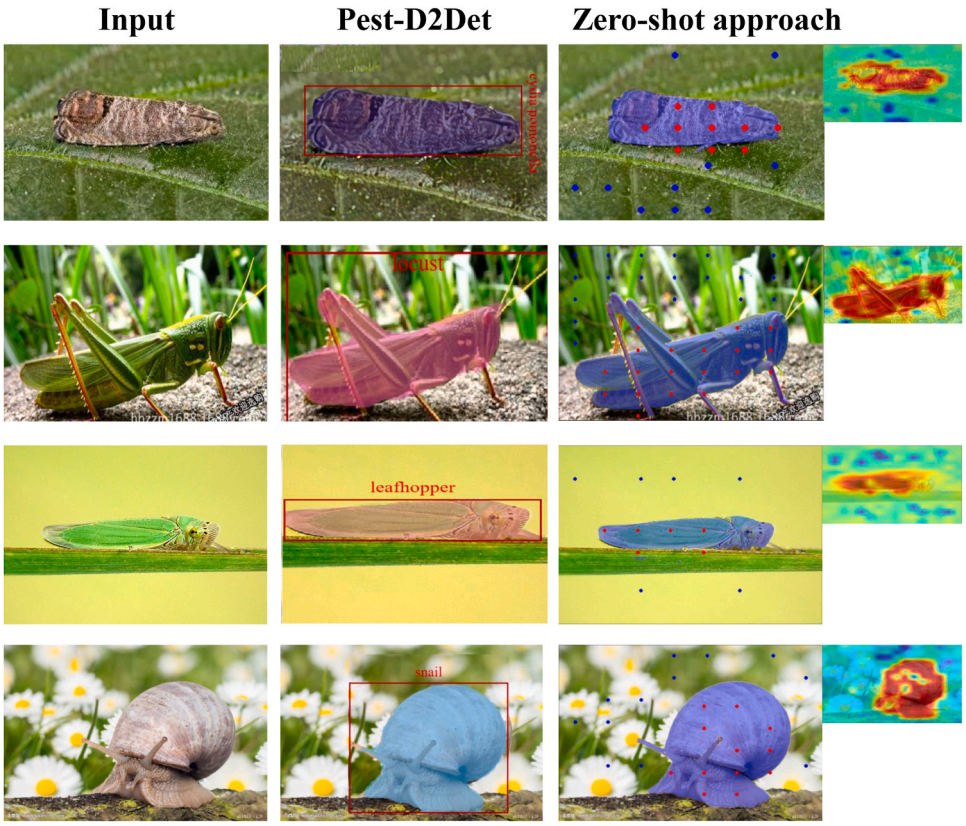

Fig. 7. Comparison of segmentation results of the Pest-D2Det model and the proposed zero-shot pest segmentation framework. **Note:** In each image of the third column, red dots indicate regions identified by the model as potential pest locations, while blue dots represent background areas. This dual visualization scheme helps guide the SAM model.

provides a more reliable performance evaluation.

Pest-D2Det demonstrates consistent performance across all five folds and outperforms other segmentation models with the highest mean mIoU of 75.22 % and a standard deviation of 1.16. SegFormer achieves strong performance with a mean mIoU of 73.8 %, but with slightly higher variability (standard deviation of 1.33). DeepLabv3 offers the most consistent performance with the lowest standard deviation of 0.90, but at the expense of lower overall performance, with a mean mIoU of 67.88 %. Finally, the zero-shot approach demonstrates reasonable stability and competitive performance compared to the supervised approaches with a mean mIoU of 65.96 %. The lower performance is due to the trade-off of not requiring extensive annotated training data.

Figure 7 presents a comparative analysis of segmentation results between the supervised Pest-D2Det model and the proposed zero-shot pest segmentation framework. Pest-D2Det was selected for segmentation comparison with the zero-shot approach because it is a specialized model designed for pest segmentation, and it demonstrated the best performance among the tested supervised models (Section 5.3.3). The first column displays the input images of various pests. The second column demon states the segmentation output of the Pest-D2Det model, where the detected pests are highlighted with red bounding boxes and masks. The third column illustrates the output of the zero-shot approach with the attention map visualization.

Overall, the zero-shot approach can identify pests in diverse and complex backgrounds, as evidenced by the attention maps on the rightmost side of each row. These attention maps demonstrate the model's focus on relevant regions, which is crucial for accurate pest detection. Occasionally, the zero-shot approach shows more precise segmentation results compared to the Pest-D2Det model. For example, in the image of the leafhopper (third row), the zero-shot model's segmentation mask accurately segments even the leg part of the leafhopper, whereas the Pest-D2Det model fails to segment some parts of the leg. However, in most cases, the zero-shot approach displays less precise boundaries between the pests and the background compared to the Pest-D2Det model. Therefore, in the future, careful fine-tuning of the zero-shot approach can be performed to improve its boundary precision.

These observations highlight the trade-offs between the two approaches. The zero-shot method offers flexibility and the ability to operate without extensive labeled data, but at the cost of segmentation precision. In contrast, the supervised model, which relies on annotated training data, provides more accurate and sharply defined pest boundaries.

6. Discussion

The original dataset used in this study was primarily collected from online platforms, typically consisting of low-resolution images that can potentially impact the model's performance. To address this issue, this study integrates the SMSR super-resolution algorithm to enhance image quality, preserve fine details, and enrich dataset diversity. The experimental results in Section 5.3.1 show that the SR module increases the average mIoU performance by 2.3 % for both supervised and zero-shot approaches, with the zero-shot approach benefiting the most (mIoU increased from 62.6 % to 66.5 %). These improvements can be attributed to several factors. Firstly, SR enhances image clarity by recovering fine details often lost during image acquisition or compression. This enables models to more accurately detect and segment pests, especially those with complex visual characteristics. Secondly, SR can mitigate the effects of noise and artifacts present in real-world images, which can interfere with the model's ability to correctly identify and localize pests. While the pre-processing phase demands additional processing time and computational power, its implementation can be easily customized based on specific application requirements.

This study investigated the effectiveness of a zero-shot approach for pest segmentation. The proposed model's performance was evaluated against three state-of-the-art supervised segmentation models, including Pest-D2Det, SegFormer, and DeepLabv3, using a manually annotated pest testing dataset. Table 3 revealed a trade-off between accuracy and processing speed. While the zero-shot approach achieved a competitive mIoU score of 66.5 %, it exhibited slightly lower performance compared to the supervised models. Moreover, this was accompanied by the lowest inference speed, which is likely influenced by the processing requirements of both ECLIP and SAM models, along with the post-processing steps involved in generating the final segmentation masks. We also analyzed the robustness of these models using a 5-fold cross-validation, as shown in Table 4. Among the supervised segmentation approaches, Pest-D2Det demonstrated the most consistent mean mIoU at 75.22 % with a relatively low standard deviation, suggesting its robustness to data variations. The proposed zero-shot approach achieved competitive segmentation results compared to supervised methods with an overall mean mIoU of 65.96 %. These findings are consistent with observations from prior research on zero-shot approaches Pourpanah et al. (2022). Notably, the proposed zero-shot pest segmentation pipeline leveraging ECLIP and SAM offers distinct advantages over conventional supervised learning methods. The model's primary strength lies in its ability to be deployed without extensive labeled image datasets, which can be expensive and laborious to collect. In addition, it shows excellent generalization capabilities to new environments and new pest species and robustness to variations in pest appearance, such as size, shape, and color.

7. Conclusions and future works

This research introduces a novel zero-shot segmentation approach for pest monitoring, which has the potential to be applied to automated farm inspection applications. To evaluate the performance of the proposed zero-shot approach compared to the latest supervised segmentation methods, a comprehensive dataset containing 12,300 images of 10 common pest species was collected.

The zero-shot segmentation pipeline seamlessly combines ECLIP and SAM. ECLIP plays a crucial role in generating potential pest locations and background regions within the image as point candidates through user prompt input. These candidates are then fed into the point-based prompt encoding module of SAM for accurate pest segmentation. The zero-shot approach achieves a good mIoU of 66.5 % on the testing dataset. In addition, the extracted attention map from ECLIP provides valuable insights into the model's decision-making process by visualizing the model's attention weights. This technique highlights the relative importance of each pixel within the input image for the model's final output. By analyzing these attention weights, experts/farmers can understand which image regions contribute most significantly to the model's predictions.

While the proposed zero-shot pest monitoring framework demonstrates potential, its application in real-time scenarios requiring precise pest localization within complex natural backgrounds is limited. Although the dataset used in this study is diverse, it does not include all possible environmental conditions encountered in agricultural settings. Extreme variations in lighting, weather, and background clutter can affect model performance. Future studies should include more diverse datasets to validate the model's robustness under varying environmental conditions. The preprocessing phase and the zero-shot model require significant computational resources. This resource intensity may limit the practical deployment of the model in resource-constrained environments. Finally, the model's ability to accurately detect small body parts such as legs and antennae is currently insufficient. To address these limitations, future work will focus on collecting more diverse datasets to test the model's robustness under varying environmental conditions, optimizing the model for computational efficiency, and improving the objective function to enhance the detection of fine-grained pest details.

CRediT authorship contribution statement

Muhammad Fayaz: Data curation. **Nur Alam:** Visualization, Data curation. **Hyoung-Kyu Song:** Supervision, Funding acquisition. **Dinh Khuong Nguyen:** Software, Conceptualization. **Hyeonjoon Moon:** Writing – review & editing, Supervision. **L. Minh Dang:** Writing – original draft, Investigation. **Asma Khan:** Visualization, Formal analysis. **Sufyan Danish:** Investigation, Data curation.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the metaverse support program to nurture the best talents (IITP-2023-RS-2023–00254529) grant funded by the Korea government(MSIT) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries(IPET) through Digital Breeding Transformation Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs(MAFRA)(322063–03–1-SB010) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540).

References

Ashok, P., Jayachandran, J., Gomathi, S.S., Jayaprakasan, M., 2019. Pest detection and identification by applying color histogram and contour detection by svm model. Int. J. Eng. Adv. Technol. 8.

Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R.E., Zhu, C., 2022. Real-world single image super-resolution: A brief review. Inf. Fusion 79, 124–145.

Chen, L.C., Papandreou, G., Schroff, F., Adam, H., 2017.Rethinking atrous convolution for semantic image segmentation.arXiv preprint arXiv:1706.05587.

Chopda, J., Raveshiya, H., Nakum, S., Nakrani, V., 2018. Cotton crop disease detection using decision tree classifier. 2018 International Conference on Smart City and Emerging Technology (ICSCET), IEEE, pp. 1–5.

Contributors, M., 2020.Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark.

Dang, M., Wang, H., Li, Y., Nguyen, T.H., Tightiz, L., Xuan-Mung, N., Nguyen, T.N., 2024. Computer vision for plant disease recognition: A comprehensive review. Bot. Rev. 1–61.

Domingues, T., Brandão, T., Ferreira, J.C., 2022. Machine learning for detection and prediction of crop diseases and pests: A comprehensive survey. Agriculture 12, 1350.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020.An image is worth 16x16 words: Transformers for image recognition at scale.arXiv preprint arXiv: 2010.11929.

Du, K., Huang, J., Wang, W., Zeng, Y., Li, X., Zhao, F., 2024.Monitoring low-temperature stress in winter wheat using tropomi solar-induced chlorophyll fluorescence.IEEE Transactions on Geoscience and Remote Sensing.

Ebrahimi, M., Khoshtaghaza, M.H., Minaei, S., Jamshidi, B., 2017. Vision-based pest detection based on svm classification method. Comput. Electron. Agric. 137, 52–58.

Jalayer, S., Sharifi, A., Abbasi-Moghadam, D., Tariq, A., Qin, S., 2023. Assessment of spatiotemporal characteristic of droughts using in situ and remote sensing-based drought indices. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16, 1483–1502.

Kasinathan, T., Uyyala, S.R., 2021. Machine learning ensemble with image processing for pest identification and classification in field crops. Neural Comput. Appl. 33, 7491–7504.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023.Segment anything.arXiv preprint arXiv: 2304.02643.

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.

Liu, J., Wang, X., 2021. Plant diseases and pests detection based on deep learning: a review. Plant Methods 17, 1–18.

Liu, L., Wang, R., Xie, C., Yang, P., Wang, F., Sudirman, S., Liu, W., 2019. Pestnet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. IEEE Access 7, 45301–45312.

Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N., 2022. Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 1–66.

Nguyen, L.Q., Shin, J., Ryu, S., Dang, L.M., Park, H.Y., Lee, O.N., Moon, H., 2023. Innovative cucumber phenotyping: A smartphone-based and data-labeling-free model. Electronics 12, 4775.

Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C.P., Wang, X.Z., Wu, Q.J., 2022.A review of generalized zero-shot learning methods.IEEE transactions on pattern analysis and machine intelligence.

Preti, M., Verheggen, F., Angeli, S., 2021. Insect pest monitoring with camera-equipped traps: Strengths and limitations. J. Pest Sci. 94, 203–217.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021.Learning transferable visual models from natural language supervision, In: International conference on machine learning, PMLR.8748-8763.

Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T., 2008. Labelme: a database and web-based tool for image annotation. Int. J. Comput. Vis. 77, 157–173.

Rustia, D.J.A., Lin, C.E., Chung, J.Y., Zhuang, Y.J., Hsu, J.C., Lin, T.T., 2020. Application of an image and environmental sensor network for automated greenhouse insect pest monitoring. J. Asia-Pac. Entomol. 23, 17–28.

Rustia, D.J.A., Lu, C.Y., Chao, J.J., Wu, Y.F., Chung, J.Y., Hsu, J.C., Lin, T.T., 2021. Online semi-supervised learning applied to an automated insect pest monitoring system. Biosyst. Eng. 208, 28–44.

Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T., 2023.Pic2word: Mapping pictures to words for zero-shot composed image retrieval, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 19305-19314.

Sciarretta, A., Calabrese, P., 2019. Development of automated devices for the monitoring of insect pests. Curr. Agric. Res. J. 7.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.

Skendžić, S., Zovko, M., Živković, I.P., Lešić, V., Lemić, D., 2021. The impact of climate change on agricultural insect pests. Insects 12, 440.

Wang, H., Li, Y., Dang, L.M., Moon, H., 2022. An efficient attention module for instance segmentation network in pest monitoring. Comput. Electron. Agric. 195, 106853.

Wang, H., Nguyen, T.H., Nguyen, T.N., Dang, M., 2024. Pd-tr: End-to-end plant diseases detection using a transformer. Comput. Electron. Agric. 224, 109123.

Wang, L., Dong, X., Wang, Y., Ying, X., Lin, Z., An, W., Guo, Y., 2021.Exploring sparsity in image super-resolution for efficient inference, In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 4917-4926.

Wang, M., Wang, B., Zhang, R., Wu, Z., Xiao, X., 2023. Flexible vis/nir wireless sensing system for banana monitoring. Food Qual. Saf. 7, fyad025.

Wang, X., Huang, J., Feng, Q., Yin, D., 2020. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. Remote Sens. 12, 1744.

Wei, Y., Cao, Y., Zhang, Z., Peng, H., Yao, Z., Xie, Z., Hu, H., Guo, B., 2023.iclip: Bridging image classification and contrastive language-image pre-training for visual recognition, In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2776-2786.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Zhong, F., Chen, Z., Zhang, Y., Xia, F., 2020. Zero-and few-shot learning for diseases recognition of Citrus aurantium l. using conditional adversarial autoencoders. Comput. Electron. Agric. 179, 105828.

Zhu, L., Wu, M., Wan, X., Zhao, N., Xiong, W., 2017. Image recognition of rapeseed pests based on random forest classifier. Int. J. Inf. Technol. Web Eng. (IJITWE) 12, 1–10.