horticulturae

MDPI

Article

Universal Image Segmentation with Arbitrary Granularity for Efficient Pest Monitoring

L. Minh Dang [1,2,3], Sufyan Danish [4], Muhammad Fayaz [4], Asma Khan [4], Gul E. Arzu [4], Lilia Tightiz [4], Hyoung-Kyu Song [3] and Hyeonjoon Moon [4,*]

1 Institute of Research and Development, Duy Tan University, Da Nang 550000, Vietnam
2 Faculty of Information Technology, Duy Tan University, Da Nang 550000, Vietnam
3 Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul 05006, Republic of Korea; songhk@sejong.ac.kr
4 Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea; sufyandanish@sju.ac.kr (S.D.); muhammadfayaz@sju.ac.kr (M.F.); asmakhan28@sju.ac.kr (A.K.); arzurabani@sju.ac.kr (G.E.A.); liliatightiz@sejong.ac.kr (L.T.)
* Correspondence: hmoon@sejong.ac.kr

Abstract

Accurate and timely pest monitoring is essential for sustainable agriculture and effective crop protection. While recent deep learning-based pest recognition systems have significantly improved accuracy, they are typically trained for fixed label sets and narrowly defined tasks. In this paper, we present RefPestSeg, a universal, language-promptable segmentation model specifically designed for pest monitoring. RefPestSeg can segment targets at any semantic level, such as species, genus, life stage, or damage type, conditioned on flexible natural language instructions. The model adopts a symmetric architecture with self-attention and cross-attention mechanisms to tightly align visual features with language embeddings in a unified feature space. To further enhance performance in challenging field conditions, we integrate an optimized super-resolution module to improve image quality and employ diverse data augmentation strategies to enrich the training distribution. A lightweight postprocessing step refines segmentation masks by suppressing highly overlapping regions and removing noise blobs introduced by cluttered backgrounds. Extensive experiments on a challenging pest dataset show that RefPestSeg achieves an Intersection over Union (IoU) of 69.08 while maintaining robustness in real-world scenarios. By enabling language-guided pest segmentation, RefPestSeg advances toward more intelligent, adaptable monitoring systems that can respond to real-time agricultural demands without costly model retraining.

Keywords: deep learning; universal segmentation; crop pest; segmentation; super resolution

check for updates

Academic Editors: Adélia Sousa and Fabricio Macedo

Received: 23 October 2025
Revised: 21 November 2025
Accepted: 2 December 2025
Published: 3 December 2025

Citation: Dang, L.M.; Danish, S.; Fayaz, M.; Khan, A.; Arzu, G.E.; Tightiz, L.; Song, H.-K.; Moon, H. Universal Image Segmentation with Arbitrary Granularity for Efficient Pest Monitoring. *Horticulturae* **2025**, *11*, 1462. https://doi.org/10.3390/horticulturae11121462

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Pest infestations have long threatened agricultural productivity by reducing crop yield and quality. Timely and effective pest management is thus critical to minimizing economic losses and ensuring sustainable production. However, traditional inspection methods, which rely on manual visual identification by farmers, are often inefficient in large-scale or complex agricultural settings [1]. This inefficiency is exacerbated by the high variability in pest morphology (e.g., size, shape) and adaptive camouflage, which impede accurate detection.

Modern agriculture increasingly utilizes drone- and camera-acquired imagery to build large-scale pest datasets [2]. Yet these datasets suffer from inconsistent image quality due to variable lighting conditions and resolution fluctuations in real-world environments. Super-resolution (SR) techniques can mitigate such issues by enhancing image clarity and detail [3], while data augmentation improves model robustness, particularly for detecting tiny or partially occluded pests.

Deep learning (DL) has recently emerged as a powerful tool for pest identification, offering superior recognition and localization capabilities [4]. Among advanced computer vision methods, referring image segmentation (RIS) dynamically segments objects using natural language descriptions [5]. RIS addresses the limitations of predefined class labels in semantic or instance segmentation. By integrating vision and language understanding, RIS enables context-aware segmentation in complex scenes.

Our contributions to this study are as follows:

- RefPestSeg, a universal pest segmentation model that unifies visual features with textual guidance for precise pest localization.
- An enhanced SR pipeline to significantly enhance image quality for precise pest analysis.
- A refined postprocessing module that resolves overlapping segmentation masks and suppresses noise.

The paper is structured as follows: Section 2 reviews related work. Section 3 provides a detailed explanation of the main components of the RefPestSeg framework. In Section 4, we describe the large-scale pest dataset and evaluation metrics. Section 5 presents experimental results and analysis. Section 6 discusses implications and limitations. Finally, Section 7 concludes the study and outlines future research directions.

2. Related Work

2.1. Pest Segmentation and Detection in Agriculture

DL has transformed agricultural pest monitoring through pixel-level segmentation. Semantic pest segmentation architectures, including Fully Convolutional Networks (FCNs) [6], U-Net variants [7], and DeepLabV3+ [8], provide dense pest localization but struggle with overlapping instances. Instance pest segmentation frameworks like Mask R-CNN and its derivatives [9,10] address this limitation by jointly detecting pests and generating pixel-precise masks, significantly improving handling of clustered pests.

Recent work further refines these baseline architectures with attention mechanisms. For example, Wang et al. [4] proposed an Efficient Channel and Spatial Attention Network (ECSA-Net) for instance pest segmentation, achieving mAP of 78.6% for pest detection and 77.2% for pest segmentation, particularly improving performance on small and camouflaged pests. In another line of work, Zhang et al. [11] enhanced pest boundary recognition under low-contrast conditions via multi-scale dilated attention. For real-time deployment, lightweight pest segmentation architectures like TinySegFormer [12] and slice-attention GRUs [13] enable field deployment of pest detection systems, with the latter reporting 99.52% accuracy and 99.1% IoU on controlled datasets.

Despite these advances, current methods remain constrained by fixed-category paradigms. They require exhaustive retraining when encountering novel pest species, morphological variants, or shifting environmental conditions. This bottleneck highlights the need for category-agnostic segmentation approaches that adapt to dynamic field requirements without architectural reconfiguration.

2.2. Language-Guided and Referring Image Segmentation

Referring image segmentation (RIS) overcomes the fixed-category limitations of traditional segmentation by enabling language-guided object localization. Unlike seman-

tic/instance segmentation, which requires predefined class labels, RIS segments arbitrary objects specified through natural language expressions. This flexibility is critical for agricultural settings where novel pest species, morphological variants, or rare infestations evade rigid category systems.

Early RIS methods [14,15] processed visual and linguistic features independently before late-stage fusion, resulting in weak cross-modal alignment and error propagation from isolated feature extractors. Subsequent work improved linguistic grounding through syntactic parsing [16,17], but struggled with ambiguous descriptions. Attention-based fusion now dominates state-of-the-art RIS. Models like LAVT [18] and PolyFormer [19] dynamically weight image regions using textual queries.

Recent advances further optimize multi-modal alignment: EAVL [20] generates dynamic convolution kernels conditioned on text inputs. RISAM [21] integrates the Segment Anything Model (SAM) with bidirectional Vision-Guided Attention and Language-Guided Attention, achieving strong zero-shot generalization.

3. Methodology

Figure 1 illustrates the RefPestSeg framework pipeline. The workflow begins with image restoration: input images undergo preprocessing via enhanced super-resolution (SR) module (Section 3.1), which recovers high-frequency details critical for identifying tiny pests. During training, data augmentation techniques are applied to enhance the diversity of the preprocessed training set. After that, the augmented data is fed into RefPestSeg model, which involves key components such as text and image encoding (Section 3.2.1), pre-fusion (Section 3.2.2), visual–linguistic decoding (Section 3.2.3), and postprocessing (Section 3.2.4). Quantitative validation using standard segmentation metrics is presented in Section 4.4.

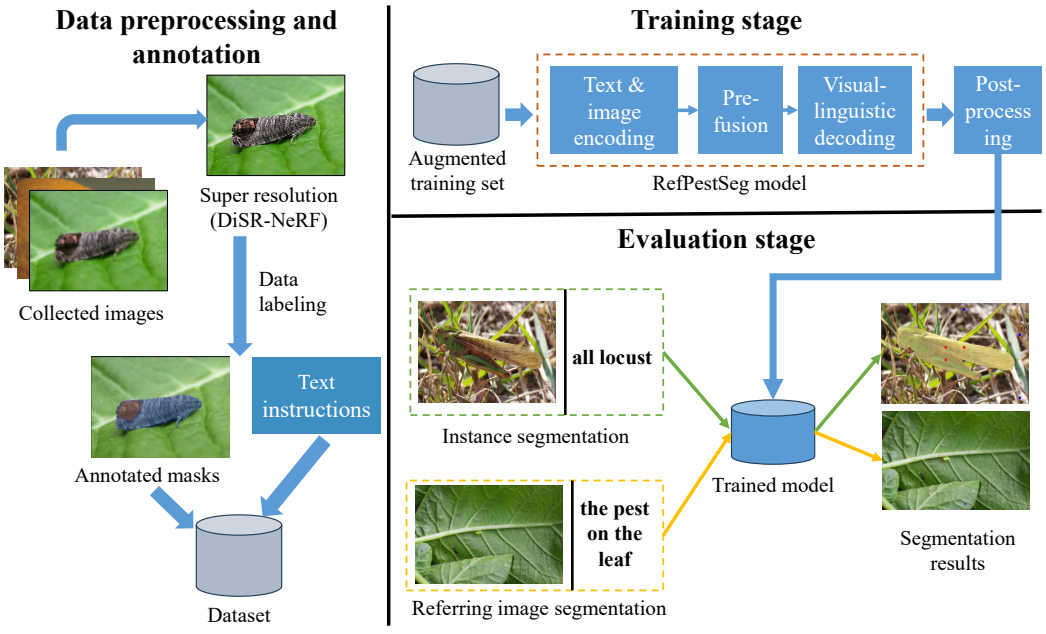


Figure 1. Diagram of the proposed RIS pest recognition framework.

3.1. Data Preprocessing

Super-resolution (SR) is essential for agricultural pest recognition where field-captured images often suffer from motion blur, sensor noise, and resolution limitations [22]. Diffusion-based SR methods have have recently shown promising results [23]. In this work, we adopt DiSR-NeRF [24], a diffusion-guided SR framework that enhances Neural Radiance Fields (NeRFs) with high-resolution, view-consistent details. DiSR-NeRF

outperforms existing methods in both image quality and multi-view consistency without requiring high-resolution training data.

Unlike conventional approaches that depend on high-resolution reference images, DiSR-NeRF leverages state-of-the-art 2D diffusion-based super-resolution models to upscale low-resolution NeRF-rendered outputs. The framework then introduces Iterative 3D Synchronization (I3DS), which alternates between (i) upscaling low-resolution NeRF renders via diffusion models, and (ii) refining the underlying 3D NeRF representation using these enhanced views. This closed-loop optimization ensures geometrically consistent high-frequency details across novel viewpoints.

To further improve sharpness and cross-view coherence, DiSR-NeRF proposes Renoised Score Distillation (RSD), a latent-space optimization technique that fuses ancestral sampling and Score Distillation Sampling (SDS). RSD minimizes the discrepancy between optimized and predicted latents at denoising step $t-1$ and can be formulated as follows.

$$L_{RSD} = \|z_{t-1} - \hat{z}_{t-1}\| \tag{1}$$

Here, z'_{t-1} is a renoised latent derived from a detail-enhanced initial latent $z'_0 = z_0 + h_\theta$ (where h_θ encodes high-frequency residuals). This latent is generated by reapplying the forward diffusion process to z'_0 at timestep t. The prediction $\hat{z}_{t-1}$ is produced by the diffusion U-Net conditioned on the noisier latent z'_t, the original low-resolution input, and timestep information. By operating directly on intermediate latents rather than pixel-space gradients, RSD enables efficient gradient flow through the full denoising trajectory, preserving input fidelity while synthesizing sharp, view-consistent details.

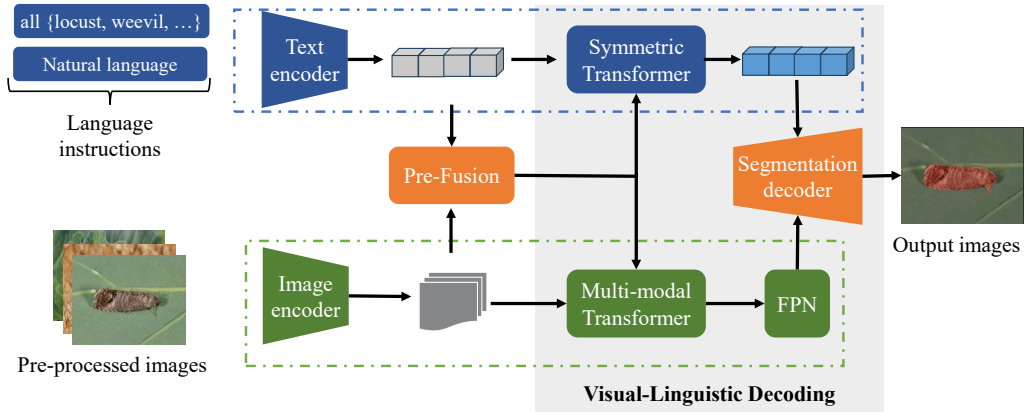
3.2. RefPestSeg Model

Figure 2a illustrates the pipeline of RefPestSeg for pest monitoring. Its core design philosophy leverages extensive visual–linguistic interactions within a unified language-guided paradigm [25]. In particular, RefPestSeg accepts both an input image and a corresponding language prompt as inputs.

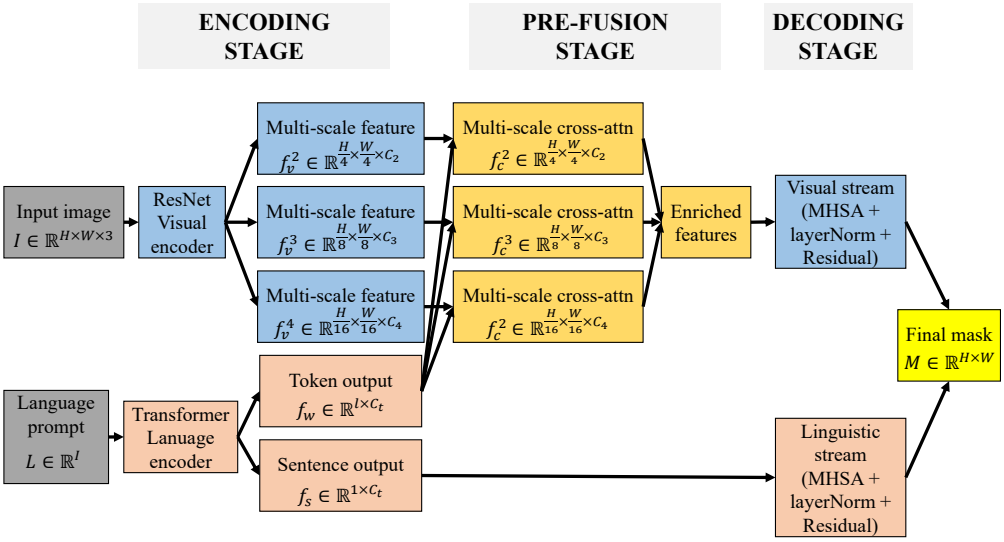
The detailed architecture of RefPestSeg (Figure 2b) comprises three sequential stages:

- Encoding Stage: Parallel visual and linguistic features are extracted. A ResNet visual encoder generates multi-scale hierarchical features at resolutions from the input image: $f_v^2 \in \mathbb{R}^{H/4\times W/4\times C_2}$, $f_v^3 \in \mathbb{R}^{H/8\times W/8\times C_3}$, $f_v^4 \in \mathbb{R}^{H/16\times W/16\times C_4}$. Concurrently, a transformer-based language encoder processes the prompt to yield token-level embeddings $f_w \in \mathbb{R}^{l\times C_t}$ and global sentence-level embedding $f_s \in \mathbb{R}^{1\times C_t}$.
- Pre-fusion Stage: Multi-scale cross-attention mechanisms align visual and linguistic features at each resolution. This fuses semantic context from f_w into the visual features, producing enriched representations f_c^2, f_c^3, and f_c^4 that preserve spatial dimensions while integrating visual structure and linguistic semantics.
- Decoding Stage: A symmetric dual-stream decoder processes the fused features. The visual stream refines enriched features f_c^k through multi-head self-attention (MHSA) blocks with layer normalization and residual connections. The linguistic stream dynamically refines text embeddings f_s via cross-attention with visual features, followed by identical MHSA-normalization blocks. The streams' outputs are fused to generate the segmentation mask $M \in \mathbb{R}^{H\times W}$.

By operating in a unified cross-modal representation space, RefPestSeg dynamically perceives segmentation targets specified by language prompts, enabling flexible segmentation at arbitrary semantic granularities, from species-level to instance-specific pest identification.



(**a**) Visualization of the RefPestSeg prediction process



(**b**) RefPestSeg feature dimensions and transformations

Figure 2. Overview of the RefPestSeg pest recognition model based on language instructions. It includes two main subfigures: (**a**) Visualization of the RefPestSeg prediction process, and (**b**) RefPestSeg feature dimensions and transformations.

3.2.1. Encoding Process

As illustrated in Figure 5, RefPestSeg receives an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a language prompt L (token sequence of length l).

- Visual Encoding: A ResNet backbone extracts multi-scale hierarchical features from I. We denote the features at three resolution levels as $f_v^i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$ for $i \in \{2,3,4\}$. H^i and W^i correspond to the spatial dimensions of the feature maps at the i-th scale, and C_v^i represents the number of channels in the extracted feature.
- Linguistic Encoding: A transformer-based language encoder processes the tokenized prompt L to generate (i) token-level embeddings $f_w \in \mathbb{R}^{l\times C_t}$ capturing word semantics and syntactic relationships; and (ii) sentence-level embedding $f_s \in \mathbb{R}^{1\times C_t}$ representing global semantic meaning. Both embeddings share the same channel dimension C_t for cross-modal compatibility.

3.2.2. Pre-Fusion Process

The pre-fusion process integrates linguistic semantics into multi-scale visual features through cross-modal attention. As shown in Figure 2b, it takes token-level embeddings f_w

and hierarchical visual features f_v^i ($i \in \{2, 3, 4\}$) as inputs. These visual features capture complementary details: high-resolution scales ($i = 2$) preserve fine textures, while deeper scales ($i = 4$) encode high-level semantics. By injecting linguistic guidance at multiple resolutions, pre-fusion enhances attention to language-relevant regions while suppressing irrelevant background noise. Following [25], we adopt a lightweight design that achieves effective feature activation without excessive complexity.

For each scale i, we compute cross-attention where visual features serve as queries and linguistic embeddings provide keys/values. The visual feature $f_v^i \in \mathbb{R}^{H_i \times W_i \times C_i}$ (with $H_i = H/2^i$, $W_i = W/2^i$) is reshaped to $\mathbb{R}^{N_i \times C_i}$ ($N_i = H_i W_i$), then projected to query space. The interaction between these modalities is mathematically formulated as follows:

$$f_c^i = softmax\left(\frac{G_q(f_v^i)^T G_k(f_w)}{\sqrt{C^i}}\right) G_v(f_w)^T \tag{2}$$

where G_q, G_k, and G_v are projection functions that map the input features into a shared embedding space. The term $\sqrt{C^i}$ is used to normalize the similarity scores for numerical stability. The output feature f_c^i is an enriched representation that encodes visual structure and linguistic semantics. f_c^i facilitates downstream visual–linguistic tasks, including referring expression comprehension, multi-modal segmentation, and attention-driven object detection.

3.2.3. Visual–Linguistic Decoding

To maximize vision–language synergy, a dual-stream decoding architecture was adopted to hierarchically integrate multi-scale image features with context-aware textual representations via multi-modal transformers [26]. This design bridges the semantic gap between modalities and yields contextually grounded predictions.

Visual stream processing: For the i-th level visual feature map $f_c^i \in \mathbb{R}^{\frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times C_i}$ from a ResNet backbone, spatial dimensions are flattened to form token sequences:

$$f_v^i = \text{flatten}(f_c^i) \in \mathbb{R}^{N_i \times C_i} \tag{3}$$

where $N_i = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$ denotes the number of spatial tokens at scale i. Spatial relationships are preserved through learnable positional embeddings $P_i \in \mathbb{R}^{N_i \times C_i}$:

$$f_v^i = f_v^i + \text{Pos.} \in \mathbb{R}^{N_i \times C_i} \tag{4}$$

To enable effective concatenation with linguistic embeddings, dimensional compatibility needs to be ensured. The visual tokens f_v^i have dimension C_i (determined by the ResNet backbone at scale i), while word embeddings $f_w \in \mathbb{R}^{l \times C_t}$ have dimension C_t (determined by the language encoder). Since C_i may differ from C_t, a learnable linear projection is applied to align visual features with the linguistic embedding space:

$$\tilde{f}_v^i = W_v f_v^i \in \mathbb{R}^{N_i \times C_t} \tag{5}$$

where $W_v \in \mathbb{R}^{C_i \times C_t}$ is a learnable projection matrix that transforms visual features from dimension C_i to C_t, enabling them to reside in the same embedding space as linguistic features.

Multi-modal fusion: Projected visual tokens $\tilde{f}_v^i$ are concatenated with word embeddings $f_w \in \mathbb{R}^{l \times C_t}$ (where l is text length) along the sequence dimension:

$$f_m^i = \text{Concat}(\tilde{f}_v^i, f_w). \tag{6}$$

The fused representation f_m^i enables the model to capture cross-modal dependencies through self-attention mechanisms, enabling it to effectively learn meaningful correlations between visual and linguistic embeddings. To further refine these fused features, the model

employs MHSA. This technique computes attention across different subspaces of the feature representation. The MHSA operation utilizes learnable weights H^i, W^i, which allow the self-attention mechanism to adaptively adjust its focus on spatial and semantic dimensions:

$$f_b^i = MHSA(f_m^i) \cdot (H^i W^i). \tag{7}$$

The computed features f_b^i are then normalized using Layer Normalization, with a residual connection ensuring stable gradient propagation:

$$f_b^i = LN(f_b^i) + f_c^i. \tag{8}$$

Linguistic stream refinement: Inspired by prompt learning [27], a language instruction updating strategy is adopted to refine the linguistic embedding space by integrating visual context. Similar to the visual stream, the attention operation is implemented to achieve this goal. The linguistic stream mirrors the multi-modal transformer used in the visual stream, resulting in a symmetric transformer architecture.

The process begins with cross-attention, where the sentence-level textual embedding f_s serves as the Query, and the activated visual features f_c act as both the Key and Value. Next, a self-attention mechanism is employed to merge the original language prompt with its context-enhanced version. Finally, the enriched features from both modalities are merged through matrix similarity computations, which enhance their alignment for subsequent tasks such as object detection and segmentation. To further refine these merged features, bilinear interpolation and thresholding techniques are applied. This process results in an accurate response mask that effectively highlights regions corresponding to the given text prompt.

3.2.4. Mask Postprocessing

Pest segmentation masks generated by the framework may contain artifacts including duplicate detections and background noise due to complex agricultural scenes. To refine these predictions, a postprocessing pipeline (Algorithm 1) is introduced to refine the segmentation results.

Algorithm 1 Postprocessing algorithm

Require: MaskList, IoU_{th}, Overlap_{th}, min_area, max_area
Step 1: Area-based mask filtering
FilteredMasks $\leftarrow \emptyset$
for each mask $m \in$ MaskList **do**
  if min_area $\leq \text{area}(m) \leq$ max_area **then**
    Add m to FilteredMasks
  end if
end for
Step 2: Iterative mask merging
for each pair (m_i, m_j) in FilteredMasks **do**
  Compute $\text{IoU}_{ij} = \frac{|m_i \cap m_j|}{|m_i \cup m_j|}$
  Compute $\text{Overlap}_{ij} = \frac{|m_i \cap m_j|}{\min(\text{area}(m_i), \text{area}(m_j))}$
  if $\text{IoU}_{ij} \geq \text{IoU}_{th}$ or $\text{Overlap}_{ij} \geq \text{Overlap}_{th}$ **then**
    $m_{\text{new}} \leftarrow \text{Merge}(m_i, m_j)$
    Replace m_i and m_j with m_{new} in FilteredMasks
  end if
end for
Step 3: Boundary refinement
for each mask $m \in$ FilteredMasks **do**
  $m \leftarrow \text{Smooth}(m)$
end for
return FilteredMasks as **FinalMasks**

The pipeline operates in three stages: (1) masks are filtered by normalized area ($\alpha_{\min} = 0.01$, $\alpha_{\max} = 0.65$) to exclude fragments and oversized regions; (2) overlapping masks are iteratively merged when their IoU exceeds $\tau_{\text{IoU}} = 0.5$ or their containment ratio (intersection over smaller mask area) exceeds $\tau_{\text{ov}} = 0.8$; and (3) final masks undergo boundary smoothing via Gaussian blurring ($\sigma = 1.0$) followed by morphological closing with a 5×5 kernel to eliminate jagged edges. This yields spatially coherent pest regions with minimal false positives.

4. Dataset and Evaluation Metric

4.1. Data Overview

This study introduces a novel pest dataset comprising 2400 images curated from two sources. First, 2000 images were manually selected from the established crop pest database of Wang et al. [4], originally captured outdoors via mobile devices and sourced from public repositories. This subset covers ten pest species: cabbage white butterfly (*Pieris rapae*), cotton leafworm (*Spodoptera littoralis*), codling moth (*Cydia pomonella*), fruit fly (*Bactrocera dorsalis*), leafhoppers (Cicadellidae family), locust (*Locusta migratoria*), mole cricket (*Gryllotalpa gryllotalpa*), snail (*Helix aspersa*), stink bug (*Halyomorpha halys*), and weevils (Curculionidae family). Second, 400 additional images of black soldier flies (*Hermetia illucens*) and crickets (Gryllidae family) were manually collected from verified online sources to enhance taxonomic diversity.

The dataset includes twelve pest categories with resolutions ranging from 224×90 to 4000×2337 pixels. The dataset was randomly partitioned into training, validation, and testing sets using an 8:1:1 ratio, generating 1920 training images, 240 validation images, and 240 test images. Representative samples for all twelve pest categories are shown in Figure 3.

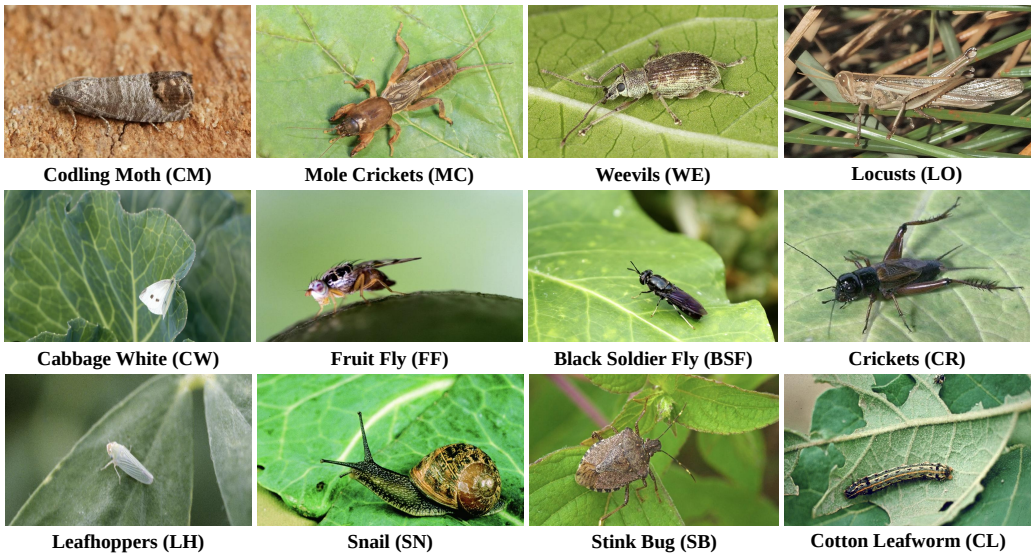


Figure 3. Representative images of the twelve pest species in the dataset. Scale varies across images due to original capture conditions.

4.2. Unified Annotation Protocol

We adopt a unified annotation format comprising (1) original RGB images, (2) pixel-accurate instance segmentation masks, and (3) task-specific language captions. Annotations were generated using LabelMe [28] following COCO standards.

4.2.1. RGB Image Constraints

Original resolutions vary significantly (224 × 90 to 4000 × 2337 pixels). All images are resized to 480 × 480 pixels. No upper resolution limit exists in the raw data. To ensure annotation reliability and practical detectability, pests occupying fewer than 32 × 32 pixels after resizing to 480 × 480 are excluded from annotation. Severely blurred or occluded specimens below this threshold are considered out-of-scope.

Images contain both single-species (multiple instances of one pest category may appear) and multi-species scenes (multiple pest species co-occur in the same image). There is no constraint that only one species or only one individual must be present in the raw image. All such configurations are allowed and are treated uniformly.

4.2.2. Image Annotation Protocol

- Each visible pest instance receives a separate polygonal mask enclosing its complete body, including head, thorax, abdomen, wings, legs, etc.
- Background elements (foliage, soil, non-pest objects) are treated as a single background class without annotation.
- Each mask is paired with a species-level categorical label.

4.2.3. Task-Specific Language Caption Annotation Protocol

To support multi-task learning while avoiding semantic conflicts [25], task-specific captions are generated as follows.

- Semantic segmentation: Template-based prompts ("all {species}") using category names as concise textual expressions.
- RIS segmentation: Instance-specific descriptions generated via BLIP [29], conditioned on pest category and visual context. This approach enables "arbitrary-granularity segmentation" at inference time, where natural language queries (e.g., "the leftmost locust" or "wings of the fruit fly") dynamically define target regions. Crucially, while part-level queries (e.g., "wings") are supported during inference, no part-level ground truth exists in the annotations. Predictions for such queries are constrained to lie within full-instance masks.

4.3. Data Augmentation

A variety of data augmentation techniques were performed on the training set to increase its diversity and robustness in real-world environments. As illustrated in Figure 4, the augmentation operations include brightness adjustment, cutout, channel shuffling, flipping, gamma contrast modification, rotation, Gaussian blur, translation, and Gaussian noise injection, each designed to simulate the natural variations present in pest images.

After augmentation, the training set increased by nine-fold to a total of 17,280 images. The substantial expansion of the training set helps reduce overfitting by exposing the model to a broader range of real-world variations, including diverse lighting conditions, occlusions, object orientations, and environmental distortions.

4.4. Evaluation Metrics

Following established protocols [30], two complementary metrics are employed to evaluate segmentation performance: Intersection over Union (IoU) and Precision@X. IoU quantifies pixel-level alignment between the predicted mask P and ground truth mask G for individual images:

$$\text{IoU} = \frac{|P \cap G|}{|P \cup G|}, \tag{9}$$

where $|P \cap G|$ and $|P \cup G|$ represent intersection and union areas, respectively. This metric provides a comprehensive measure of segmentation accuracy, with higher values indicating better spatial correspondence.

Precision@X evaluates localization capability at the image level by measuring the percentage of test samples achieving IoU that exceeds a specified threshold X:

$$\text{Precision@}X = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(\text{IoU}_i \geq X) \times 100\%, \tag{10}$$

where N denotes the total test images, IoU_i is the per-image IoU score, and $\mathbb{I}(\cdot)$ is the indicator function that outputs 1 when the condition is met and 0 otherwise. Results for thresholds $X \in \{0.5, 0.7, 0.9\}$ are reported during evaliation to assess robustness under progressively stringent overlap requirements. Higher thresholds emphasize stricter boundary precision.

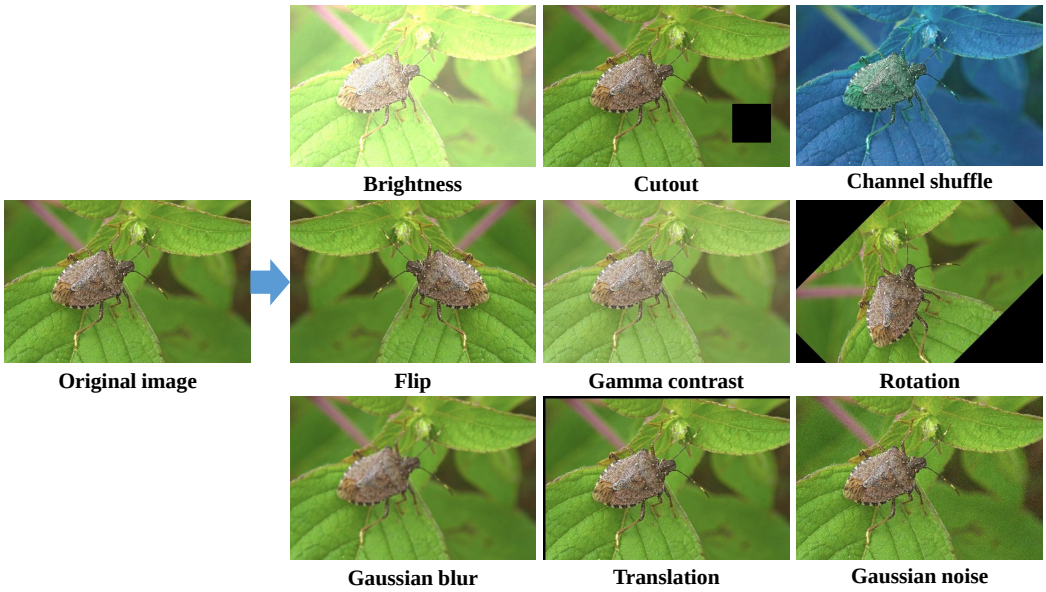


Figure 4. Sample outputs after applying different augmentation techniques on the original image.

5. Experimental Results

This section presents a comprehensive evaluation of RefPestSeg on the collected pest dataset. Section 5.1 describes the experimental setup, including hardware configurations, software environments, and hyperparameter settings for both RefPestSeg and comparative models. Section 5.2 systematically assesses model capabilities through four key analyses.

- Preprocessing impact (Section 5.2.1): Quantifies accuracy gains from the preprocessing pipeline.
- Segmentation performance (Section 5.2.2): Reports quantitative metrics and visual results across diverse scenarios.
- Granularity analysis (Section 5.2.3): Evaluates segmentation fidelity at multiple semantic levels (low to high granularity).
- Comparative benchmarking (Section 5.2.5): Validates RefPestSeg's performance against state-of-the-art segmentation baselines.

5.1. Implementation Descriptions

RefPestSeg is initialized with ImageNet-pretrained weights. CLIP ViT-B/16 [31] is adopted as the text encoder and ResNet-50 [32] for the vision backbone. Input images are resized to 480 × 480 pixels during training and inference.

The model was trained on an NVIDIA Tesla V100 GPU with a batch size of 32 for 15 epochs. We used the Adam optimizer with an initial learning rate of 1×10^{-4}, decayed by a factor of 0.1 after epoch 10. To stabilize feature extraction, the visual backbone employed a reduced learning rate scaled by 0.1 relative to other components. The loss function combined Dice loss and cross-entropy loss with equal weighting [33]. For textual grounding, we randomly sampled one natural language description per pest category per training iteration.

During inference, predicted masks were upsampled to original resolution via bilinear interpolation and binarized at 0.5 to generate binary segmentations. No postprocessing was applied. For instance, for segmentation tasks, language prompts followed the template "`all [pest category]`" (e.g., "all locust").

5.2. Pest Identification Framework Performance Assessment

5.2.1. Performance Analysis of the Preprocessing Process

Image preprocessing is critical for enhancing the quality of the pest dataset collected from various sources, and it involves applying the DiSR-NeRF super-resolution model [24] to upscale the images by a scaling factor of $\times 4$. The performance of DiSR-NeRF is quantitatively and qualitatively evaluated using the Peak Signal-to-Noise Ratio (PSNR), which objectively measures the improvement in image quality.

Figure 5 illustrates representative results for locust and snail specimens. Regions of interest (ROIs) are compared across ground truth high-resolution (HR) images, LR inputs, and DiSR-NeRF outputs. The reconstructed images exhibit markedly sharper edges and finer texture details than LR inputs. Specifically, the locust's segmented body and appendages gain visible definition, while the snail's shell patterns and contours become easily recognizable. DiSR-NeRF achieves high-fidelity reconstruction with PSNR exceeding 32, demonstrating its effectiveness in preserving biologically critical pest features. This enhancement directly benefits downstream tasks such as pest identification and segmentation.

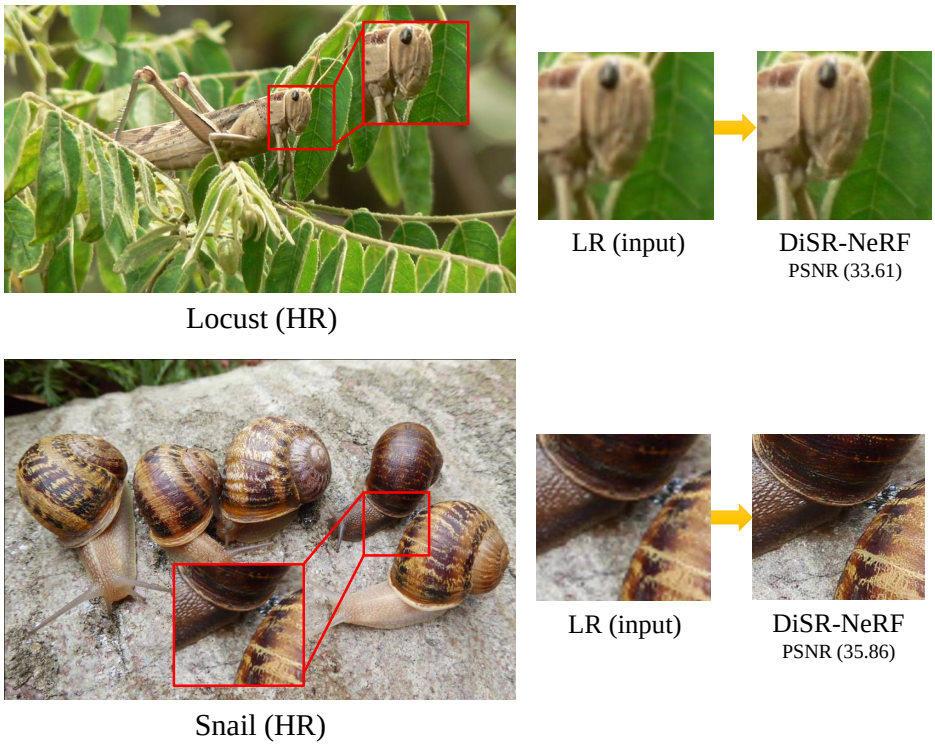


Figure 5. Super-resolution results using DiSR-NeRF ($\times 4$ scaling). Note: SR refers to the super-resolution output, LR denotes low-resolution inputs, and PSNR stands for Peak Signal-to-Noise Ratio.

Quantitative validation (Table 1) confirms that DiSR-NeRF preprocessing substantially boosts segmentation performance for the RefPestSeg model. Without preprocessing, RefPestSeg achieves an IoU of 64.50 and precision scores of 73.22, 68.25, and 58.10 at confidence thresholds of 0.5, 0.7, and 0.9, respectively. With DiSR-NeRF preprocessing, IoU increases to 69.08, while precision improves to 77.94, 73.31, and 63.67 at the same thresholds. These gains highlight the module's ability to enhance boundary delineation and fine-detail recovery, particularly crucial for small or partially occluded pests, confirming its robustness in real-world agricultural settings.

Table 1. Impact of DiSR-NeRF preprocessing on RefPestSeg segmentation performance.

	IoU	**Precision@0.5**	**Precision@0.7**	**Precision@0.9**
w/o preprocessing	64.50	73.22	68.25	58.10
w/ preprocessing	69.08	77.94	73.31	63.67

5.2.2. RefPestSeg Performance Evaluations

A comprehensive evaluation of RefPestSeg was conducted across three progressively challenging scenarios: (1) standard instance segmentation for single pests, (2) instance segmentation for multiple co-occurring pests, and (3) language-guided RIS. Qualitative results are visualized in Figure 6. The first two rows demonstrate instance segmentation capabilities, while the third row illustrates RIS results guided by textual prompts (e.g., "the pest in the middle," "the wings of the fruit fly").

Figure 6. Qualitative segmentation results of RefPestSeg. (**Top two rows**): Instance segmentation for single and multiple pests. (**Bottom row**): Referring image segmentation (RIS) guided by textual prompts. Masks are color-coded by instance (IS) or language-guided RIS.

RefPestSeg consistently generates precise masks that capture morphological details (e.g., wings, body segments) and maintain clear boundaries even under challenging conditions like partial occlusions and color variations similar to backgrounds. Crucially, the RIS functionality transcends conventional category-based segmentation by dynamically adapting to linguistic instructions. This enables targeted analysis of arbitrary regions of interest (e.g., specific body parts), establishing RefPestSeg as a versatile tool that bridges traditional instance segmentation and context-aware agricultural pest analysis.

Quantitative evaluation on the pest dataset (Table 2) shows RefPestSeg achieves a mean IoU of 69.08% across 12 pest categories. Highest performance occurs on morphologically distinct pests with clear boundaries: snail (75.42% IoU) and codling moth (73.19%). Strong results are also observed for crickets (72.35%) and cotton leafworm (71.84%), indicating effective handling of well-defined structures. Performance decreases for pests with irregular shapes or high background similarity: mole crickets (63.18%) and fruit fly (64.57%), where segmentation is challenged by structural complexity and camouflage effects.

Table 2. Per-category segmentation performance of RefPestSeg. Metrics: Intersection over Union (IoU) and precision at confidence thresholds (Pr@*k*).

Pest	IoU	Pr@50	Pr@70	Pr@90
Black Soldier Fly	70.25	80.43	75.12	65.37
Mole Crickets	63.18	69.87	64.53	55.29
Weevils	68.49	78.36	73.84	63.57
Locusts	65.22	72.19	68.63	60.15
Cabbage White	66.78	74.62	70.55	61.38
Snail	75.42	85.12	80.97	70.25
Codling Moth	73.19	83.47	78.84	68.32
Crickets	72.35	82.76	77.93	67.54
Leafhoppers	67.89	75.62	71.48	62.97
Fruit Fly	64.57	71.93	67.28	58.46
Stink Bug	69.74	79.28	74.15	64.02
Cotton Leafworm	71.84	81.59	76.43	66.78
Average	69.08	77.94	73.31	63.67

Precision analysis at increasing confidence thresholds (Pr@*k*) reveals robustness to false positives at moderate thresholds (Pr@50: 77.94% average), with snail (85.12%) and codling moth (83.47%) showing highest reliability. Performance naturally declines at stringent thresholds (Pr@90: 63.67% average), particularly for structurally complex pests like mole crickets (Δ = 14.58%) and fruit fly (Δ = 13.47%), reflecting inherent challenges in high-confidence segmentation of ambiguous features.

However, as shown in Table 2, increasing the decision threshold from Pr@50 to Pr@90 results in a substantial precision drop (from 77.94% to 63.67% on average), with especially significant drops for morphologically complex or camouflaged species such as mole crickets and fruit fly. This suggests that under stringent confidence requirements, the model either under-segments ambiguous regions or rejects otherwise correct predictions. Despite these challenges, RefPestSeg demonstrates strong adaptability for a wide range of pest species. The model's ability to maintain consistent performance on diverse morphological structures makes it a promising tool for precision agriculture and automated pest monitoring systems.

The segmentation results in Figure 7 highlight the effectiveness of the proposed model in handling three challenging scenarios: tiny pest recognition, occlusion, and noises. In the first case, the model successfully recognizes a tiny pest despite its small size and low contrast on a complex leaf background. The second scenario illustrates the robust segmentation performance of RefPestSeg on a partially occluded cricket, where the model successfully differentiates the pest from its surroundings. Lastly, in environments with severely noisy backgrounds, the model effectively differentiates the pest from cluttered

backgrounds and generates a mask that retains critical pest features while minimizing false positives. These results demonstrate the model's ability to precisely detect small or partially occluded pests in complex environments by isolating relevant fine-grained details of pests from background noise.

Figure 7. Robustness demonstration of RefPestSeg in handling challenging cases including tiny pests on complex backgrounds, partially occluded specimens, and noisy environments with distractors.

5.2.3. Analysis of RefPestSeg Segmentation Results at Various Semantic Levels

Figure 8 demonstrates RefPestSeg's capability to interpret linguistic instructions across semantic hierarchies. These results illustrate a progressive understanding of the model, ranging from low to high semantic complexity.

At the lowest semantic level, the model achieves precise part-based segmentation, accurately isolating anatomical structures (e.g., "head" of locusts and crickets). Moving to object-level segmentation, it successfully identifies entire pest instances (e.g., fruit flies and black soldier flies), demonstrating robust instance discrimination through language guidance. Crucially, RefPestSeg handles both singular and plural queries (e.g., "all locust heads"), confirming its versatility in multi-instance scenarios, a critical requirement for agricultural pest monitoring.

At higher semantic levels, the model transitions to class-based and contextual understanding. It accurately segments entire pest categories (e.g., "all weevils" or "all codling moths"), distinguishing between taxonomically similar species. At the highest level, it processes scene-level descriptions like "fly, leaf," simultaneously segmenting pests and relevant environmental elements. This contextual awareness, which integrates background semantics with target objects, enables practical deployment in complex agricultural settings where pest–environment interactions are diagnostically significant.

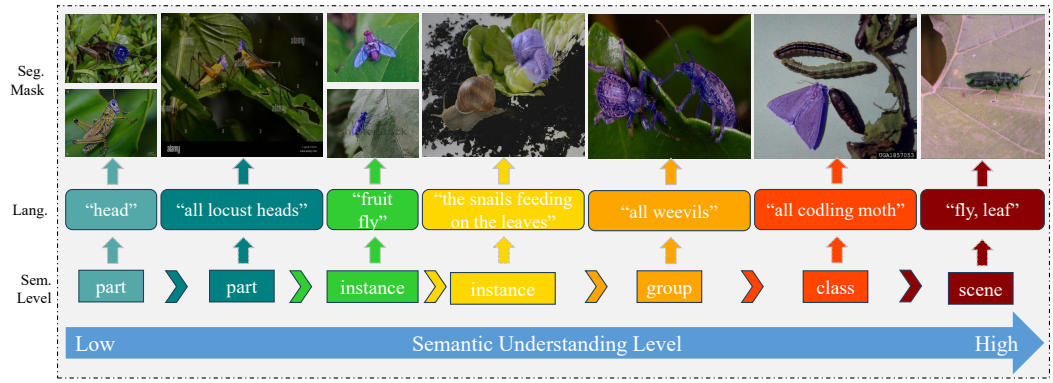


Figure 8. RefPestSeg segmentation results for different semantic levels guided by language instructions. Note: "Seg. Mask", "Lang.", and "Sem. Level" denote segmentation masks, language descriptions, and semantic levels (part → instance → class → scene), respectively.

5.2.4. Ablation Study

To verify that the architecture of RefPestSeg is well aligned with the language prompt paradigm, an ablation experiment on the pre-fusion module (PF), vision path (VP), and language path (LP) is conducted, and the results are summarized in Table 3. Starting from the baseline, introducing PF alone improves IoU from 58.42 to 61.85 and Precision@50 from 67.31 to 70.48. This indicates the benefit of early cross-modal interaction. Adding either the VP or LP on top of PF further boosts performance. The addition of both paths simultaneously achieves the best results of 69.08 IoU and 77.94 Precision@50. Overall, these ablations demonstrate that strengthening the interaction between visual and language streams consistently enhances RefPestSeg on the collected dataset.

Table 3. Ablation studies about the model components.

Method	IoU	Precision@50
Baseline	58.42	67.31
+Pre-fusion	61.85	70.48
+Pre-fusion, + Vision path	65.23	74.12
+Pre-fusion, + Language path	64.97	73.56
+Pre-fusion, + Vision path, + Language path	69.08	77.94

5.2.5. Comparison Study for RefPestSeg

Table 4 compares RefPestSeg against state-of-the-art segmentation models on pest segmentation performance. While RefPestSeg achieves competitive results across multiple metrics, its unique value lies in combining visual segmentation with language-driven flexibility, a capability absent in conventional approaches.

Accuracy trade-offs: RefPestSeg achieves an IoU of 69.08%, positioning it between FCN (66.35%) and Mask R-CNN (70.10%). Its IoU is lower than specialized architectures like Pest-D2Det (75.41%) and transformer-based models like SegFormer (74.25%), suggesting opportunities for improvement in boundary refinement and multi-scale feature integration.

Precision–speed balance: RefPestSeg achieves the second-highest Precision@50 (77.94%) among all models, outperforming DeepLabv3 (75.12%) and SegFormer (77.35%), while operating at 16 FPS. This exceeds the inference speed of Pest-D2Det (12 FPS) and SegFormer (11 FPS) by around 30%. While FCN (22 FPS) and DeepLabv3 (18 FPS) offer higher throughput, their substantially lower precision ($\geq$9.74% gap) makes them less suitable for precision-critical applications.

Practical significance: RefPestSeg uniquely considers accuracy, efficiency, and interactive adaptability. Its language-guided segmentation capability enables dynamic adaptation

to novel pest descriptions without retraining, a critical advantage over static visual-only models. This positions RefPestSeg as an optimal solution for precision agriculture systems requiring: (1) near real-time performance (16 FPS), (2) high segmentation fidelity (Precision@50 77%), and (3) contextual understanding through natural language queries.

Table 4. Performance metrics of various segmentation models on the pest dataset1.

Model	IoU	Precision@50	FPS
FCN [34]	66.35	68.20	22
Mask R-CNN [35]	70.10	72.45	14
Pest-D2Det [4]	75.41	79.43	12
DeepLabv3 [36]	72.85	75.12	18
Segformer [37]	74.25	77.35	11
RefPestSeg (Ours)	69.08	77.94	16

6. Discussion

The RefPestSeg model integrates an advanced image preprocessing pipeline to enhance the quality of input data prior to segmentation. Leveraging the DiSR-NeRF super-resolution model during the preprocessing stage, images are upscaled by a factor of 4×, significantly improving fine-grained details and clarity. As demonstrated in the quantitative and qualitative evaluations presented in Section 5.2.1, the preprocessing module plays a critical role in enhancing segmentation robustness, particularly for pests that are small, partially occluded, or embedded in cluttered backgrounds. By extracting finer edge details and morphological features, the model ensures more accurate mask generation for downstream applications such as automated pest monitoring and precision agriculture.

RefPestSeg demonstrates exceptional segmentation capabilities under various conditions (Figure 6). The model effectively distinguishes individual pest instances and accurately identifies multiple pests in complex, cluttered backgrounds. Unlike traditional segmentation methods that rely on fixed, predefined categories, RefPestSeg dynamically adjusts its segmentation focus based on semantic-level queries. Therefore, it allows users to target specific pest body parts, individual pests, or entire species categories with remarkable precision. By bridging the gap between conventional instance segmentation and interactive, language-guided approaches, RefPestSeg facilitates targeted pest analysis and enhances agricultural decision-making efficiency. Its flexibility and adaptability make it a powerful tool for addressing the increasing challenges of pest management.

However, the present work uses a single, globally fixed threshold across all pest categories and scenes, without investigating species-specific or context-adaptive calibration. In real deployments, threshold selection will likely need to be tuned to the risk tolerance and operational constraints of each use case, and future work should explore uncertainty-aware calibration and adaptive thresholding strategies to stabilize performance at high confidence levels.

7. Conclusions and Future Works

The RefPestSeg model introduces a novel, language-guided approach to pest segmentation by combining high-resolution preprocessing with adaptive segmentation techniques. The integration of DiSR-NeRF super-resolution into the preprocessing module significantly enhances fine-grained feature extraction and boundary detection for pests. The experimental results demonstrate that the preprocessing module effectively improves segmentation accuracy, with notable improvements in IoU and precision across various confidence thresholds. Moreover, RefPestSeg outperforms several conventional segmentation models in segmentation performance while maintaining competitive computational efficiency. Its ability to dynamically adjust segmentation granularity based on text-based instructions

makes it a versatile and intelligent system, particularly suited for real-world agricultural applications such as automated pest monitoring.

RefPestSeg overcomes the limitations of traditional instance segmentation by incorporating semantic-level instructions. The model excels in recognizing tiny pests, handling occlusion challenges, and filtering out noisy backgrounds. In addition, its capacity to differentiate between specific pest body parts, individual instances, and entire pest species categories enhances its context-awareness and practical utility in precision agriculture.

Author Contributions: Conceptualization, A.K. and G.E.A.; methodology, L.M.D. and L.T.; investigation, A.K. and G.E.A.; writing—original draft preparation, L.M.D.; writing—review and editing, S.D. and M.F.; visualization, L.M.D. and S.D.; supervision, H.M.; funding acquisition, H.-K.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP), under the metaverse support program to nurture the best talents (IITP-2024-RS-2023-00254529), grant funded by the Korea government (MSIT) and by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620) and by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Technology Commercialization Support Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA)(RS-2025-02218444).

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to privacy.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Passias, A.; Tsakalos, K.A.; Rigogiannis, N.; Voglitsis, D.; Papanikolaou, N.; Michalopoulou, M.; Broufas, G.; Sirakoulis, G.C. Insect Pest Trap Development and DL-Based Pest Detection: A Comprehensive Review. *IEEE Trans. Agrifood Electron.* **2024**, *2*, 323–334.
2. Wang, R.; Liu, L.; Xie, C.; Yang, P.; Li, R.; Zhou, M. Agripest: A large-scale domain-specific benchmark dataset for practical agricultural pest detection in the wild. *Sensors* **2021**, *21*, 1601.
3. Su, H.; Li, Y.; Xu, Y.; Fu, X.; Liu, S. A review of deep-learning-based super-resolution: From methods to applications. *Pattern Recognit.* **2024**, *157*, 110935.
4. Wang, H.; Li, Y.; Dang, L.M.; Moon, H. An efficient attention module for instance segmentation network in pest monitoring. *Comput. Electron. Agric.* **2022**, *195*, 106853.
5. Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; Tan, T. Locate then segment: A strong pipeline for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9858–9867.
6. Gong, H.; Liu, T.; Luo, T.; Guo, J.; Feng, R.; Li, J.; Ma, X.; Mu, Y.; Hu, T.; Sun, Y.; et al. Based on FCN and DenseNet framework for the research of rice pest identification methods. *Agronomy* **2023**, *13*, 410.
7. Ye, W.; Lao, J.; Liu, Y.; Chang, C.C.; Zhang, Z.; Li, H.; Zhou, H. Pine pest detection using remote sensing satellite images combined with a multi-scale attention-UNet model. *Ecol. Inform.* **2022**, *72*, 101906.
8. Bose, K.; Shubham, K.; Tiwari, V.; Patel, K.S. Insect image semantic segmentation and identification using UNET and DeepLab V3+. In *ICT Infrastructure and Computing: Proceedings of ICT4SD 2022, Goa, India, 29–30 July 2022*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 703–711.
9. Li, H.; Shi, H.; Du, A.; Mao, Y.; Fan, K.; Wang, Y.; Shen, Y.; Wang, S.; Xu, X.; Tian, L.; et al. Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-RNet. *Front. Plant Sci.* **2022**, *13*, 922797.
10. Rong, M.; Wang, Z.; Ban, B.; Guo, X. Pest Identification and Counting of Yellow Plate in Field Based on Improved Mask R-CNN. *Discret. Dyn. Nat. Soc.* **2022**, *2022*, 1913577.
11. Zhang, C.; Zhang, Y.; Xu, X. Dilated inception U-Net with attention for crop pest image segmentation in real-field environment. *Smart Agric. Technol.* **2025**, *11*, 100917.

12. Zhang, Y.; Lv, C. TinySegformer: A lightweight visual segmentation model for real-time agricultural pest detection. *Comput. Electron. Agric.* **2024**, *218*, 108740.
13. Biradar, N.; Hosalli, G. Segmentation and detection of crop pests using novel U-Net with hybrid deep learning mechanism. *Pest Manag. Sci.* **2024**, *80*, 3795–3807.
14. Hu, R.; Rohrbach, M.; Darrell, T. Segmentation from natural language expressions. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 108–124.
15. Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Yuille, A. Recurrent multimodal interaction for referring image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1271–1280.
16. Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; Li, B. Referring image segmentation via cross-modal progressive comprehension. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10488–10497.
17. Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; Berg, T.L. Mattnet: Modular attention network for referring expression comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1307–1315.
18. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H. Lavt: Language-aware vision transformer for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18155–18165.
19. Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R.K.; Mahadevan, V.; Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18653–18663.
20. Yan, Y.; He, X.; Wang, W.; Chen, S.; Liu, J. Eavl: Explicitly align vision and language for referring image segmentation. *arXiv* **2023**, arXiv:2308.09779.
21. Zhang, M.; Liu, Y.; Yin, X.; Yue, H.; Yang, J. Risam: Referring image segmentation via mutual-aware attention features. *arXiv* **2023**, arXiv:2311.15727.
22. Meng, C.; Shu, L.; Han, R.; Chen, Y.; Yi, L.; Deng, D.J. Farmsr: Super-resolution in precision agriculture field production scenes. In Proceedings of the 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Beijing, China, 17–20 August 2024; pp. 1–6.
23. Moser, B.B.; Shanbhag, A.S.; Raue, F.; Frolov, S.; Palacio, S.; Dengel, A. Diffusion models, image super-resolution, and everything: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *36*, 11793–11813.
24. Lee, J.L.; Li, C.; Lee, G.H. DiSR-NeRF: Diffusion-Guided View-Consistent Super-Resolution NeRF. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 20561–20570.
25. Liu, Y.; Zhang, C.; Wang, Y.; Wang, J.; Yang, Y.; Tang, Y. Universal segmentation at arbitrary granularity with language instruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 3459–3469.
26. Zou, X.; Dou, Z.Y.; Yang, J.; Gan, Z.; Li, L.; Li, C.; Dai, X.; Behl, H.; Wang, J.; Yuan, L.; et al. Generalized decoding for pixel, image, and language. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 15116–15127.
27. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16816–16825.
28. Russell, B.C.; Torralba, A.; Murphy, K.P.; Freeman, W.T. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* **2008**, *77*, 157–173.
29. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
30. Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; Liu, T. Cris: Clip-driven referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11686–11695.
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
33. Yeung, M.; Sala, E.; Schönlieb, C.B.; Rundo, L. Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Comput. Med. Imaging Graph.* **2022**, *95*, 102026.

34. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
36. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848.
37. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.