agronomy

MDPI

Article

A Rapid Construction Method for High-Throughput Wheat Grain Instance Segmentation Dataset Using High-Resolution Images

Qi Gao [1], Heng Li [1,*], Tianyue Meng [1], Xinyuan Xu [1], Tinghui Sun [1], Liping Yin [2] and Xinyu Chai [1,3,*]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[2] Technical Center for Animal Plant and Food Inspection and Quarantine of Shanghai Customs, Shanghai 200002, China
[3] Vision Science and Rehabilitation Engineering Laboratory, Shanghai Jiao Tong University, Shanghai 200025, China
* Correspondence: liheng@sjtu.edu.cn (H.L.); xychai@sjtu.edu.cn (X.C.)

Abstract: Deep learning models can enhance the detection efficiency and accuracy of rapid on-site screening for imported grains at customs, satisfying the need for high-throughput, efficient, and intelligent operations. However, the construction of datasets, which is crucial for deep learning models, often involves significant labor and time costs. Addressing the challenges associated with establishing high-resolution instance segmentation datasets for small objects, we integrate two zero-shot models, Grounding DINO and Segment Anything model, into a dataset annotation pipeline. Furthermore, we encapsulate this pipeline into a software tool for manual calibration of mislabeled, missing, and duplicated annotations made by the models. Additionally, we propose preprocessing and postprocessing methods to improve the detection accuracy of the model and reduce the cost of subsequent manual correction. This solution is not only applicable to rapid screening for quarantine weeds, seeds, and insects at customs but can also be extended to other fields where instance segmentation is required.

Keywords: deep learning; instance segment; segment anything; annotation pipeline

Citation: Gao, Q.; Li, H.; Meng, T.; Xu, X.; Sun, T.; Yin, L.; Chai, X. A Rapid Construction Method for High-Throughput Wheat Grain Instance Segmentation Dataset Using High-Resolution Images. *Agronomy* **2024**, *14*, 1032. https://doi.org/10.3390/agronomy14051032

Academic Editor: Yanbo Huang

Received: 31 March 2024
Revised: 24 April 2024
Accepted: 9 May 2024
Published: 13 May 2024

1. Introduction

1.1. Background and Motivation

The global food crisis is a long-standing issue of food supply and demand in impoverished countries, which is triggered by international circumstances [1,2]. Global trade in food commodities is active, and the import and export of grains can easily lead to the introduction of foreign organisms, resulting in biological invasions [3,4]. To address the risks of biological invasions associated with import and export trade, countries have implemented their own customs policies. Taking China as an example, during the importation process, food grains undergo on-site inspections to ensure compliance with relevant Chinese standards and regulations [5]. If any non-compliance is found, the food grains will be refused entry into the country.

On-site testing of grains is an important step in ensuring food safety and biosecurity. However, the testing process is time-consuming and requires skilled professionals for identification. With the continuous development of global trade, the demand for imported grains is increasing, and the requirements for grain testing are becoming more stringent. Currently, customs quarantine work remains labor-intensive. Regarding pest quarantine work, the entire process, from sampling and photography to laboratory testing, is manually operated, resulting in not only time and labor inefficiency but also a need for improvement in detection rates.

This situation fails to meet the demands of high-throughput, efficiency, and intelligence. Therefore, there is a need for more intelligent and efficient detection methods. Such methods can not only improve the accuracy and reliability of testing but also reduce labor costs, expedite customs clearance, and ensure food safety.

One proposed solution for implementing rapid on-site screening of imported grains using artificial intelligence technology is as follows [6]: The selected samples are spread out and laid flat, and their images are captured using a camera. An object detection model is then utilized to locate and classify the samples in the images. To ensure quick screening, the camera's field of view should not be too small. On the other hand, to ensure that the model can identify the object categories, the pixel resolution of individual objects should not be too low [7]. Crop particles, weed seeds, and insect samples are all small in size, thus requiring a high pixel resolution for the entire image. In essence, the key characteristic of achieving rapid on-site customs inspection is the identification of small objects in ultra-high-resolution images.

However, training instance segmentation models requires the establishment of pixel-level ground truth annotations, which can be labor-intensive and time-consuming. Therefore, it is necessary to develop methods for constructing low-cost and efficient annotation datasets. Currently, both visual and language models have made rapid advancements, and large-scale zero-shot models can be used to assist in fast annotation, thereby reducing manual annotation costs. We propose a method for quickly constructing pixel-level mask annotation datasets by leveraging multiple visual models. This allows annotators to rapidly annotate object positions and masks in images using text-based annotations. Furthermore, we integrated this annotation pipeline into a visualization software tool to facilitate manual correction of the model's annotations by annotators.

In summary, we have developed a pipeline suitable for low-cost and efficient annotation of small object detection and segmentation in ultra-high-resolution images, with a focus on rapid screening of customs quarantine weed seeds and insects. This solution is also applicable to similar scenarios such as pedestrian detection in high-resolution images. Our contributions are as follows:

- We have proposed simple and effective preprocessing and postprocessing algorithms to enhance the accuracy of small object detection in ultra-high-resolution images.
- We have built a pipeline for fast ground truth annotation, which improves the efficiency of dataset construction and reduces the associated costs.
- We have packaged the annotation pipeline into software, making it convenient for the manual correction of model annotations.

1.2. Related Work

1.2.1. Weed Seed Datasets and Detection Models

Deep learning models have demonstrated significant research potential in assisting weed seed detection, with their advantage lying in the ability to automatically extract complex features from images and effectively address seed recognition issues under different environmental conditions. However, the performance of these models heavily relies on the quality and diversity of the training dataset. Currently, there have been efforts to construct relevant datasets, such as the DeepWeeds dataset [8], which covers eight weed categories and includes 17,590 labeled images, providing abundant resources for model training. However, this dataset primarily focuses on outdoor weed images, with a limited coverage of seed images, which, to some extent, restricts the application of models in seed detection tasks. Luo et al. [6] designed data collection hardware and successfully collected data for 140 weed seed species, including 33,600 instances of training data and 14,096 of testing data. Although these data offer new possibilities for weed seed detection, they have not made this dataset publicly available, limiting its application in broader research.

In addition to dataset construction, some studies also focus on model selection and optimization [9–12]. For example, Luo et al. [6] compared six popular CNN models in their research and found that AlexNet performed well in terms of classification accuracy and

efficiency, while GoogLeNet achieved optimal accuracy. These studies not only provide a rich selection of models for weed seed detection but also offer valuable references for future research.

1.2.2. Ultra-High-Resolution Small Object Detection

In the screening of weed seeds in the customs importation of grain, a wide field of view is required for rapid screening, while high resolution is necessary to ensure object recognition accuracy. Therefore, the detection of small objects in ultra-high-resolution images is particularly crucial for the rapid screening of grain. In addition to grain rapid screening, similar challenges are also present in scenarios such as remote sensing image monitoring and pedestrian detection.

Ultra-high-resolution small object detection tasks currently face two major challenges. Firstly, they require significant computational resources. Secondly, as the network depth increases, the feature information of small objects may gradually dilute or get lost during the propagation process. Existing research has proposed two main approaches. The first approach is region-based processing. Specific methods include dividing the high-resolution image into multiple smaller blocks and treating it as a problem of object detection on these smaller block images. Alternatively, an enlargement strategy can be employed, where potential object regions are detected first, followed by an enlargement process applied to these regions. Some models focus on processing only the regions that may contain objects while ignoring other regions to reduce the computational burden [13].

The second approach is the fusion of low-resolution and high-resolution features. For example, high-resolution images can be inputted into shallow networks to capture positional information, while low-resolution images can be inputted into deep networks to extract rich semantic information [14]. Another fusion strategy involves initially predicting the coarse position of small objects based on low-resolution features and then refining the detection results with the guidance of high-resolution features [15,16].

1.2.3. Applications of Grounding DINO and SAM Models

Grounding DINO and SAM, both renowned for their zero-shot capabilities, exhibit remarkable performance across various domains through prompt-based controls. This flexibility has led to their widespread utilization in specialized tasks such as lung segmentation in chest radiographs [17], in 3D with nerfs [18], semantic segmentation of artworks [19], and agriculture [20]. Recent efforts have explored the synergistic potential of combining these models. For instance, Ren et al. [21] experimented with integrating Grounding DINO and SAM to achieve text-guided image segmentation, further enhancing this framework by incorporating stable-diffusion for text-controlled image editing or OSX for promptable human motion analysis. Additionally, Jiao et al. [22] leveraged the CLIP model to categorize the segmentation outputs of SAM. Beyond direct application, these models have also been utilized as data sources [23].

However, simply concatenating these models for general tasks is insufficient for our targeted high-resolution small object detection. Direct application in this scenario would likely lead to failed predictions. Therefore, it is imperative to devise a scheme that modifies the model inputs, ensuring they align with the data distribution effectively processed by the models. This approach promises to enhance the accuracy and reliability of small object detection in high-resolution imagery.

2. Materials and Methods

We employed a combination of multiple zero-shot large models to establish a rapid annotation pipeline, as shown in Figure 1. We packaged the pipeline into software for convenient manual correction of annotation errors. Furthermore, we propose a preprocessing method to improve the zero-shot accuracy of models and reduce the time required for manual correction. To test the effectiveness of the proposed pipeline, we captured images of wheat, weeds, and insects, using a 6.5 million pixel camera as experimental subjects.

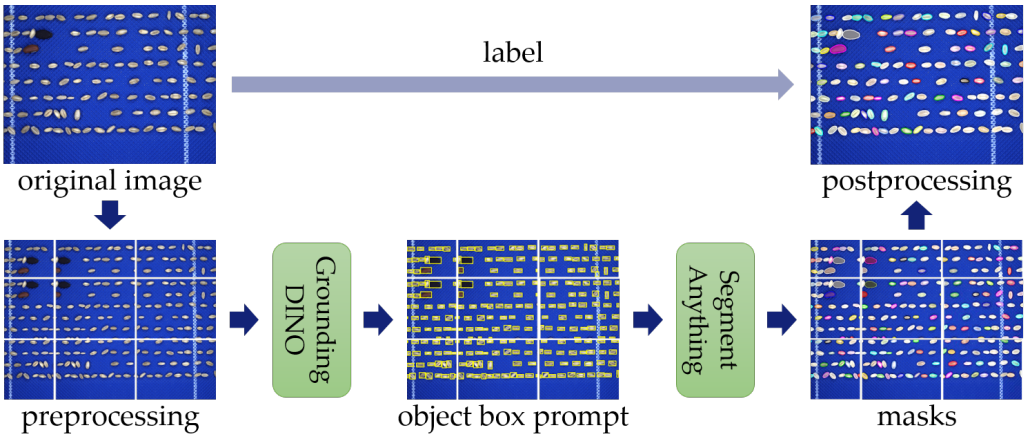


Figure 1. A pipeline for annotating ultra-high-resolution object detection datasets based on the Grounding DINO and SAM models. First, the images are divided into overlapping patches, and then, the Grounding DINO model is used to obtain bounding boxes. The bounding box results are then fed as prompts into the SAM model to generate masks. Finally, a merging algorithm is employed to integrate the annotation structures of the patches.

2.1. Capturing Images for Testing

To enhance the visibility of the foreground objects, such as grains, weed seeds, and insects, we utilized a blue background. The camera was positioned directly above the blue belt, with a field of view measuring 20 cm × 30 cm. The shooting distance was set at 20 cm. With a camera pixel count of 65 million, an object with an original size of approximately 1 mm would be represented by around 100 pixels in the captured image. The hardware equipment used for taking the photographs is illustrated in Figure 2.

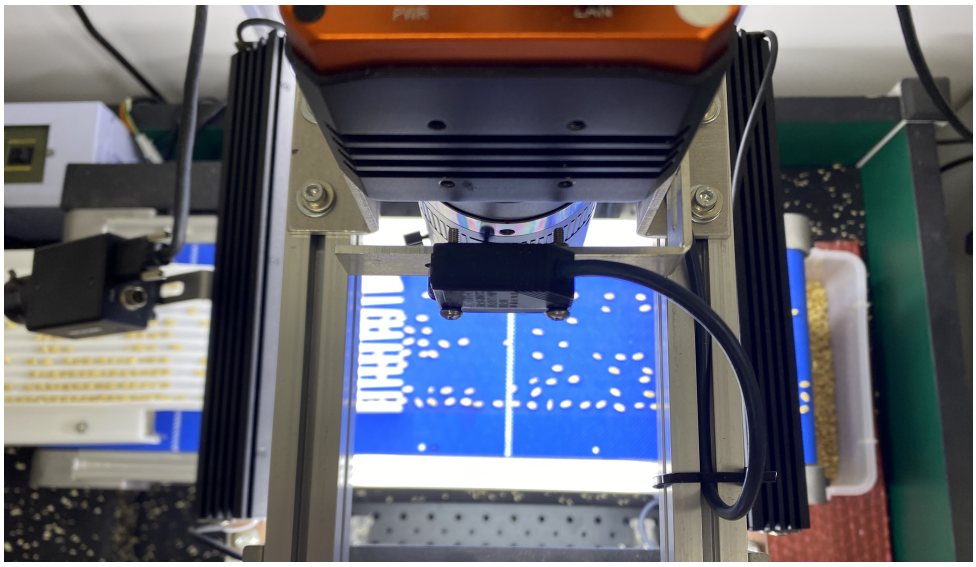


Figure 2. Hardware diagram for capturing images used in experimental testing.

2.2. Annotation Pipeline

High-throughput detection of weed seeds and insects requires both high-resolution imagery and efficient algorithms. While the use of industrial cameras for capturing high-resolution images is now a mature technology, high-resolution inputs pose a significant challenge for models. Typically, the input image size range for models is from 224 × 224 to 1024 × 1024 pixels. The pre-trained zero-shot model is unable to detect objects in a large image consisting of 65 million pixels. The input image is resized to a fixed size when sent into a model. Despite the fact that both Grounding DINO and SAM are pre-trained large

models exhibiting zero-shot capabilities, simply downsampling high-resolution images and feeding them into the model can still result in prediction failures. This underscores the necessity for zero-shot abilities to adhere to specific dataset distribution requirements. Evidently, retraining the model to broaden its supported input distribution range would be prohibitively expensive. Instead, preprocessing techniques can be employed to modify the characteristics of high-resolution images, enabling them to align with the distribution of pre-trained models. This approach is analogous to the principle underlying the "prompt" technique, facilitating the adaptability of models to diverse data inputs.

To overcome this issue, we have developed a preprocessing algorithm that slices the images into patches. Each patch is then individually fed into the model for object detection. Subsequently, we have designed a postprocessing algorithm to integrate the predictions from multiple patches and generate a comprehensive detection output.

2.2.1. Preprocessing

The original size of the image is $W \times H$. We divide the image into a grid of overlapping patches with dimensions $m \times n$. The overlap rate, denoted as θ, represents the proportion of overlapping pixels between adjacent patches in relation to the total number of pixels in the original image.

As depicted in Figure 3, not all four sides of a patch have overlapping areas. Specifically, the left and upper sides of each patch contain the overlapping area, while the patches on the leftmost and topmost sides do not have any overlapping area. The shape of the patch located at position $[0, 0]$ is $\frac{W}{m} \times \frac{H}{n}$. The shape of the patch located at position $[i, j]$ is $W(\theta + \frac{1}{m}) \times H(\theta + \frac{1}{n})$.

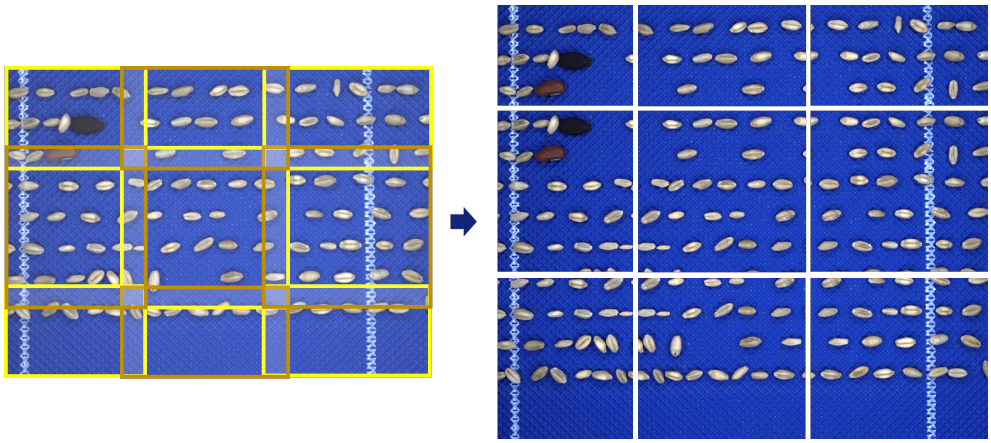


Figure 3. Image preprocessing method: splitting a high-resolution image into overlapping small patches. The yellow boxes in the image indicate the locations of the patches. The two shades of yellow are solely for ease of visualization and carry no semantic distinction.

2.2.2. Postprocessing

The split patches are inputted into the zero-shot models to generate predicted results. To prevent repeated detection of objects that appear in adjacent patches simultaneously, we propose a method for merging, as shown in Figure 4. The pseudocode is as in Algorithm 1.

Based on the relationship between the detected object's bounding box coordinates and the coordinates of the overlapping areas, we can categorize the detected objects into four types:

- Incomplete objects falling on the edges of patches;
- Objects completely falling within non-overlapping regions;
- Objects partially falling within overlapping regions;
- Objects completely falling within non-overlapping regions.

We retain objects that are completely or partially within non-overlapping regions and discard the remaining objects. This ensures that we obtain a comprehensive list of objects without any omissions or duplicates. The detailed process is shown in Figure 4.

Algorithm 1 The corresponding postprocessing method for the overlapping preprocessing method.

```
OverlapALL ← 1 // Completely in the overlap area
OverlapHALF ← 2 // Half in the overlap area
OverlapNO ← 0 // Not in the overlap area
OverlapCUT ← −1 // Cut in half by the picture, discard
Threshold ← 0.003
function TAGOBJECTS(objPos, overlapPos)
    x1, x2, y1, y2 ← objPos // Determine whether it is at the edge
    if x1 < Threshold||x2 > (1 − Threshold) then
        return OverlapCUT
    end if
    if y1 < Threshold||y2 > (1 − Threshold) then
        return OverlapCUT
    end if// Determine whether it is in the overlap area
    h, w ← overlapPos
    tag ← OverlapNO // Determine whether it is on the left side
    if x1 < w then
        if x2 < w then
            return OverlapALL
        else
            tag ← OverlapHALF
        end if
    end if// Determine whether it is on the upper side
    if y1 < h then
        if y2 < h then
            return OverlapALL
        else
            return OverlapHALF
        end if
    end if
    return tag
end function
```

2.2.3. Detection

The Grounding DINO model can receive a text prompt and output the bounding box of the detected object. The SAM model accepts prompts of the types "points" and "bounding boxes". The inputs and outputs of these two models can be concatenated to form a pipeline.

- Grounding DINO: Grounding DINO is a multimodal model that combines visual and linguistic features. By providing the object category as a text prompt, the model can output the corresponding bounding box for the object of interest. The model utilizes both the text prompt and visual features to generate the bounding box predictions. It's important to note that modifying the input text prompt can impact the model's visual feature extraction and subsequent prediction results.
- Segment Anything: Segment Anything is another multimodal model that incorporates both visual and linguistic information. This model accepts two types of prompts: points and boxes. If no specific prompt is provided, the model uses pre-defined grid points as prompts to detect all objects within the entire image.

The annotation pipeline follows the following steps:

1. The annotator designs the text prompt, specifying the desired object category or categories to be detected.
2. The text prompt and corresponding images are inputted into the Grounding DINO model, which generates bounding box coordinates for the objects based on the given prompt.
3. The original images and the obtained bounding box coordinates are then fed into the Segment Anything model. This model produces segmentation results.

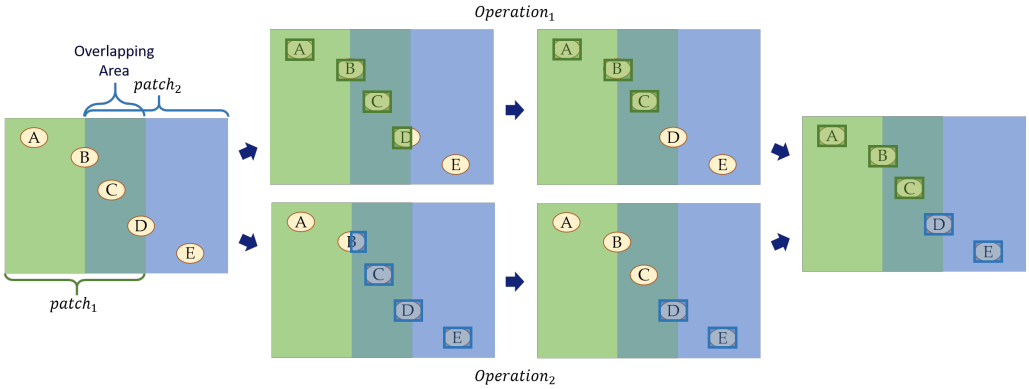


Figure 4. Merging algorithm for overlapping patches. Blue and green represent different patches. Based on the coordinates of the object's bounding boxes and the overlapping regions, objects that are incomplete and located at the edges of the image (object D in $patch_1$ and object B in $patch_2$) are identified. Additionally, objects that are entirely within the overlapping region in $patch_2$ (object C) and objects that are partially within the overlapping region (object D) are also determined. $Operation_1$ marks the incomplete object (object D) in $patch_1$ as "DELETE" while other objects (object A, B, and C) are marked as "RETAIN". $Operation_2$ labels objects in $patch_2$ that are incomplete (object B), fully (object C), or partially (object E) within the overlapping region as "DELETE", while the remaining objects are marked as "RETAIN".

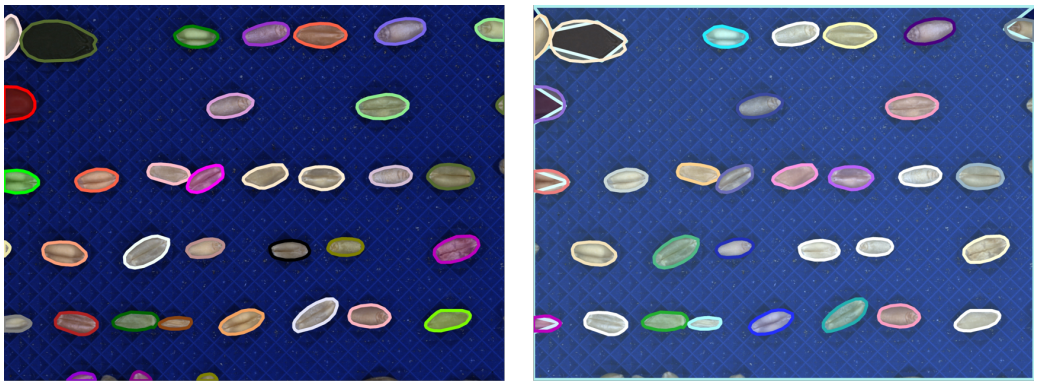
2.3. Manual Correction

Manual correction is crucial to achieve higher-quality annotations. The primary reason for this is the inherent inaccuracy of the Grounding DINO model. Being a multimodal model with complex feature fusion, altering the input text prompt can impact the model's extraction of image features. When detecting multiple object categories simultaneously, it is necessary to separate the objects using '.' in the text prompt, as per the model's settings. However, in practice, prompts containing multiple object categories often yield much worse results compared to prompts containing only one object category (as shown in Figure 5). Therefore, manual correction of the model's annotated results becomes necessary.

To enhance the efficiency of manually correcting annotation results, we have encapsulated the annotation pipeline into visual and interactive software. The annotation interface of the software is shown in Figure 6. This software needs to achieve the following functions:

- Modifying incorrect object category labels;
- Correcting inaccurate mask annotations;
- Adding mask labels for missed objects;
- Deploying the Grounding DINO model and SAM model within the software to enable users to easily achieve pixel-level annotations using text or bounding boxes.

The project and code of the software is available at https://github.com/gaoCleo/quick-label, accessed on 30 March 2024 [24].



(a) (b)

Figure 5. (**a**) Using "wheat" as the prompt. (**b**) Using "wheat . weed" as the prompt. When the number of detection categories in the prompt is increased, the accuracy of the Grounding DINO model decreases (misdetections and missed detections), resulting in a reduction in the accuracy of pipeline annotations.

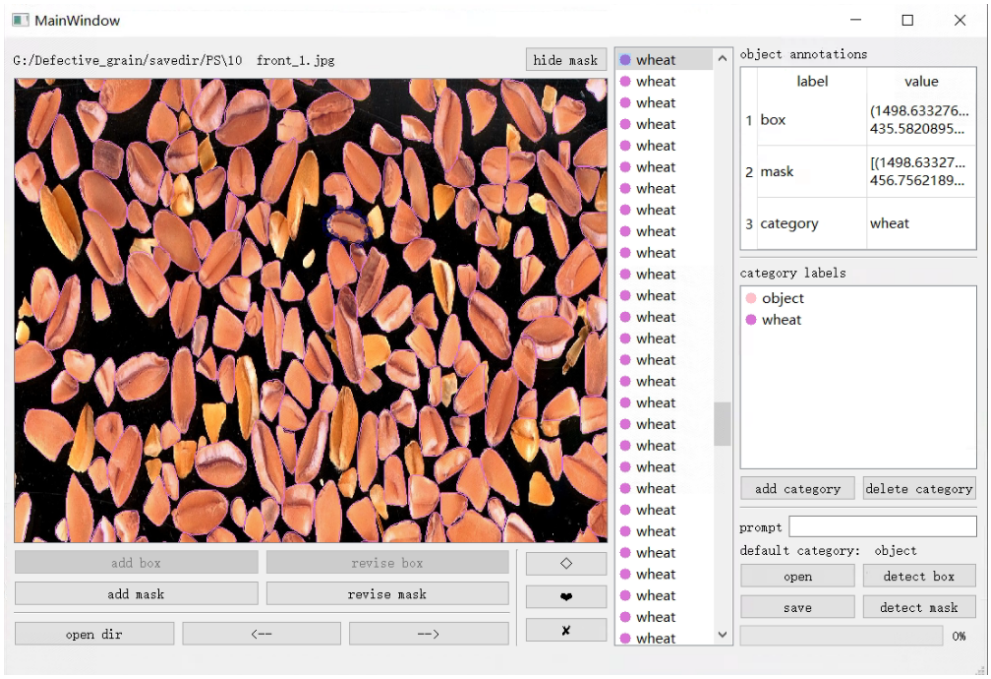


Figure 6. The annotation interface of the software.

3. Results

To demonstrate the effectiveness of our dataset construction pipeline in enhancing model annotation accuracy and reducing annotation time costs, we conducted an experiment. Specifically, we generated an ultra-high-resolution dataset comprising impurities such as wheat, diverse weed seeds, and wheat straw. Each image in the dataset had dimensions of 9000 $\times$ 7000 pixels. The primary labeling task involved identifying wheat and weed seeds, while disregarding impurities like wheat stalks. The experimental findings substantiate the superiority of our proposed pipeline in enhancing labeling accuracy and reducing time costs.

We performed two ablation experiments to validate the effectiveness of the dataset preprocessing method and the composition of the labeling pipeline, respectively. All the personnel involved in the annotation testing were skilled annotators, ensuring that the proficiency of the annotators did not impact the experimental results. Each experiment involved labeling fifty images, and we repeated each experiment five times. The data

presented in the table represent the average results across the five experiments. This experimental design enabled us to assess the impact of different factors on annotation accuracy and time costs.

We used mask AP (mask average precision) to evaluate the accuracy of various annotation methods. Since creating the dataset required highly accurate annotations, we set the mask IoU (Intersection over Union) threshold to a very high value (0.98). We considered the manually annotated ground truth, which was created using overlapped segmentation preprocessing, as the reference with an accuracy of 1.00.

3.1. Data Preprocessing and Postprocessing

To assess the effectiveness of our proposed pre-treatment and post-treatment methods, we conducted an ablation experiment in which we compared three options:

- Baseline: This option involves directly subsampling the original image and inputting it into the Grounding DINO model. The output of the Grounding DINO model is then fed into the SAM model.
- Non-overlapping clipping method: In this approach, the image is divided into $M \times N$ patches, ensuring that there is no overlap between these patches. Each patch is individually processed by the model to obtain the positioned and segmented structure. As shown in Figure 7, objects located on the dividing line between adjacent patches are merged into the same object using a synthesis algorithm.
- Overlapping clipping method: This is the method we propose. The image is segmented into overlapping patches, and a postprocessing algorithm is used to filter out redundantly labeled objects (Figure 4).

By comparing and analyzing the results of these three options, we can evaluate the effectiveness of our preprocessing and postprocessing methods in improving annotation accuracy and reducing time costs. According to the results in Table 1, the approach of directly downsampling the images without cropping and feeding them into the model resulted in a prediction accuracy of 0. In other words, the calibration process of this approach is equivalent to manually annotating the images again. We speculate that the reason for this is that aggressively downsampling the ultra-high-resolution images leads to significant loss of image information and even alters the image features, resulting in erroneous predictions by the model. Additionally, due to the large size of the images, they needed to be resized for annotation, leading to lower accuracy in mouse positioning. Therefore, even after calibration, the accuracy remains low.

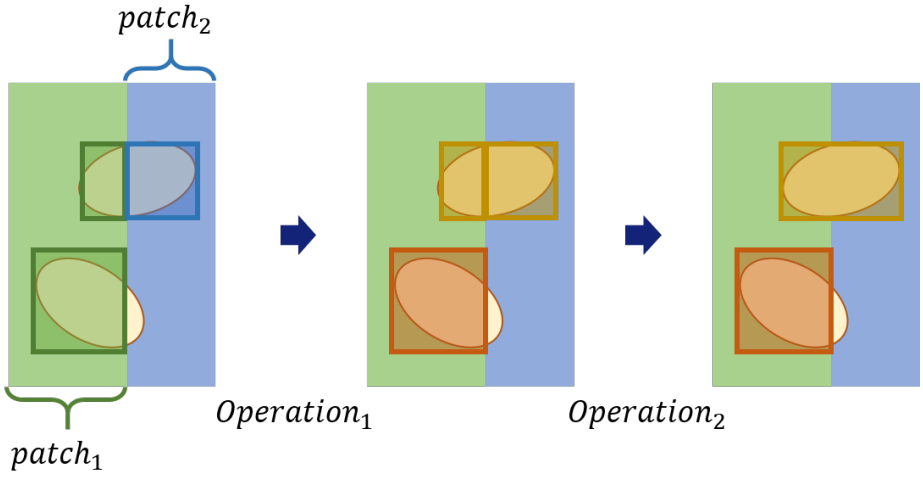


Figure 7. Merging algorithm for non-overlapping patches. Blue and green represent different patches. $Operation_1$ matches the two objects at the edges based on their coordinates. $Operation_2$ employs a closing operation to remove any gaps. However, if an object has only a minimal portion falling within $patch_2$, it is likely to be missed in $patch_2$, resulting in an incomplete mask in the final output.

Table 1. Results of ablation study on different preprocessing methods. Manual correction time refers to the total time taken to annotate 50 images.

Method	Model-Only Accuracy	Accuracy after Manual Calibration	Manual Correction Cost
Baseline	0.00	0.95	6 h 33 min
Non-overlapping	0.93	0.97	37 min
Overlapping	0.89	0.99	54 min

3.2. The Composition of the Pipeline

To enhance the prediction accuracy of zero-shot models and reduce the manual annotation time by adjusting the input distribution, we tested various pipeline combinations:

- Baseline: pure manual annotation.
- SAM-only: directly inputting the images into the SAM model.
- Manual bounding box + SAM: annotators draw bounding boxes on the images, and then the SAM model is used to predict masks.
- Grounding DINO + SAM: the images are first inputted into the Grounding DINO model and then into the SAM model.

The results, as shown in Table 2, indicate that the approach using Grounding DINO + SAM achieves the highest model prediction accuracy and requires the shortest manual calibration time.

Table 2. Results of the ablation study on pipelines with different configurations. Manual correction time refers to the total time taken to annotate 50 images.

Method	Model-Only Accuracy	Accuracy after Manual Calibration	Manual Correction Cost
Baseline	/	1.00	30 h 32 min
SAM-only	0.74	0.99	11 h 23 min
Manual box+SAM	0.99	0.99	2 h 28 min
Grounding DINO+SAM	0.89	0.99	54 min

4. Discussion

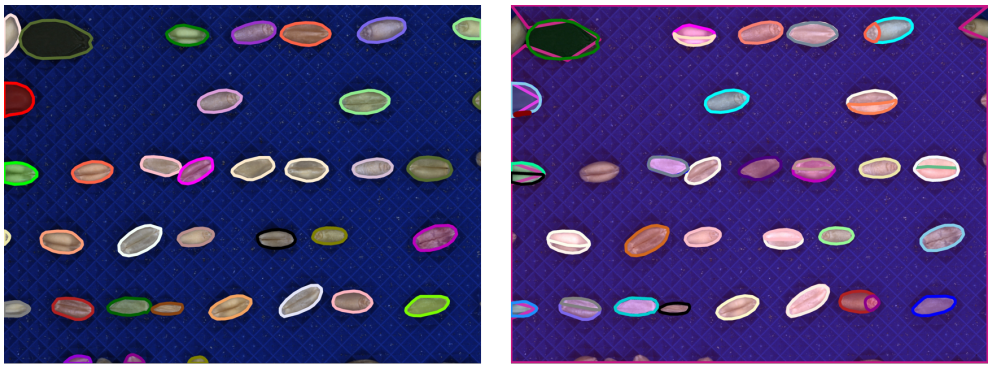
4.1. Data Preprocessing and Postprocessing

From Table 1, we can observe that the overlapping cropping strategy achieves the highest model accuracy, while the non-overlapping cropping strategy has the shortest manual calibration time. When synthesizing the results obtained from the non-overlapping cropping method, accurate annotation of objects that are split in half at the edges is required during the annotation process; otherwise, it may lead to missed detections or partial annotations instead of complete annotations. On the other hand, the overlapping cropping method avoids such issues. Hence, the accuracy of the overlapping cropping method is higher. The non-overlapping cropping produces sub-images that are consistent in size and smaller than those obtained from overlapping cropping. Consequently, the predictions from the non-overlapping cropping are more accurate, requiring fewer calibrations and less annotation time. However, the non-overlapping cropping method may have missed detections at the edges, and even after manual calibration, some omissions may still exist. This leads to incomplete or missed annotations of objects at the cropping boundaries (as shown in Figure 7), resulting in less accurate annotations with this approach.

4.2. The Composition of the Pipeline

The main issue with using SAM alone without Grounding DINO is that the SAM model's predictions contain a significant amount of over-annotation and duplicate annotations (as shown in Figure 8), and the model fails to segment objects located at the edges

of the images. Consequently, calibrating these annotations requires more time. On the other hand, Grounding DINO utilizes non-maximum suppression algorithms to suppress multiple bounding box detections on the same object, thereby mitigating the issue of over-segmentation in SAM. Correcting these errors may even take more time than manually annotating bounding boxes and then using SAM to predict masks. This further emphasizes the importance of appropriate prompts for SAM. Both the object detection and segment model components of the pipeline are interchangeable plugins. In future work, we can further compare various state-of-the-art grain instance segmentation methods [25–27] and replace the detection component of the pipeline.



(a) (b)

Figure 8. (**a**) The mask result predicted by our pipeline. (**b**) The direct prediction outcome from SAM. Direct usage of SAM often leads to issues such as redundant segmentation (where an object is segmented multiple times) and over-segmentation (where a single wheat stalk is divided into multiple parts).

5. Conclusions

We propose a solution for rapidly constructing high-resolution small object detection datasets by combining visual and language models, enabling fast pixel-level mask annotation. Our approach involves overlapping cropping preprocessing and merging postprocessing methods, which adapt ultra-high-resolution images to the input distribution of the models, thereby improving prediction accuracy. Using our proposed pipeline to annotate the dataset can save approximately 74.53% of the time and achieve nearly similar annotation results. Our proposed method and developed software not only provide technical support for the efficient construction of datasets for customs quarantine weeds and insect screening, but also have broad application prospects in dataset construction for tasks with similar requirements, such as pedestrian detection in high-resolution images and unmanned driving.

Author Contributions: Conceptualization, X.C. and H.L.; methodology, Q.G. and T.M.; software, Q.G.; validation, T.S., L.Y. and T.M.; formal analysis, X.X.; investigation, H.L.; resources, H.L. and L.Y.; data curation, X.X., T.M., T.S. and L.Y.; writing—original draft preparation, Q.G., T.M. and X.X.; writing—review and editing, Q.G., H.L., T.S., L.Y. and X.C.; visualization, X.X.; supervision, H.L. and X.C.; project administration, X.C.; funding acquisition, H.L. and X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2021YFD1400100, 2021YFD1400102); the National Natural Science Foundation of China (No. 62103269, 62073221); and the Med-X Research Fund of Shanghai Jiao Tong University (No. YG2022QN077).

Data Availability Statement: Data are available upon request from researchers who meet the eligibility criteria. Kindly contact the first author privately through e-mail.

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

References

1. Wang, X.; Ma, L.; Yan, S.; Chen, X.; Growe, A. Trade for food security: The stability of global agricultural trade networks. *Foods* **2023**, *12*, 271. [CrossRef] [PubMed]
2. Erenstein, O.; Jaleta, M.; Mottaleb, K.A.; Sonder, K.; Donovan, J.; Braun, H.J. Global trends in wheat production, consumption and trade. In *Wheat Improvement: Food Security in a Changing Climate*; Springer International Publishing: Cham, Switzerland, 2022; pp. 47–66.
3. Barratt, B.I.; Colmenarez, Y.C.; Day, M.D.; Ivey, P.; Klapwijk, J.N.; Loomans, A.J.; Mason, P.G.; Palmer, W.A.; Sankaran, K.; Zhang, F. Regulatory challenges for biological control. In *Biological Control: Global Impacts, Challenges and Future Directions of Pest Management*; CSIRO Publishing: Clayton, VIC, Austrilia, 2021; pp. 166–196.
4. Jhariya, M.K.; Banerjee, A.; Raj, A.; Meena, R.S.; Khan, N.; Kumar, S.; Bargali, S.S. Species invasion and ecological risk. In *Natural Resources Conservation and Advances for Sustainability*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 503–531.
5. Zhao, J.; Hu, K.; Chen, K.; Shi, J. Quarantine supervision of wood packaging materials (WPM) at Chinese ports of entry from 2003 to 2016. *PLoS ONE* **2021**, *16*, e0255762. [CrossRef] [PubMed]
6. Luo, T.; Zhao, J.; Gu, Y.; Zhang, S.; Qiao, X.; Tian, W.; Han, Y. Classification of weed seeds based on visual images and deep learning. *Inf. Process. Agric.* **2023**, *10*, 40–51. [CrossRef]
7. Miller, J.P.; Taori, R.; Raghunathan, A.; Sagawa, S.; Koh, P.W.; Shankar, V.; Liang, P.; Carmon, Y.; Schmidt, L. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 7721–7735.
8. Olsen, A.; Konovalov, D.A.; Philippa, B.; Ridd, P.; Wood, J.C.; Johns, J.; Banks, W.; Girgenti, B.; Kenny, O.; Whinney, J.; et al. DeepWeeds: A multiclass weed species image dataset for deep learning. *Sci. Rep.* **2019**, *9*, 2058. [CrossRef] [PubMed]
9. Sapkota, B.B.; Hu, C.; Bagavathiannan, M.V. Evaluating cross-applicability of weed detection models across different crops in similar production environments. *Front. Plant Sci.* **2022**, *13*, 837726. [CrossRef] [PubMed]
10. Peteinatos, G.G.; Reichel, P.; Karouta, J.; Andújar, D.; Gerhards, R. Weed identification in maize, sunflower, and potatoes with the aid of convolutional neural networks. *Remote Sens.* **2020**, *12*, 4185. [CrossRef]
11. Dang, F.; Chen, D.; Lu, Y.; Li, Z. YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton production systems. *Comput. Electron. Agric.* **2023**, *205*, 107655. [CrossRef]
12. Haq, M.A. CNN based automated weed detection system using UAV imagery. *Comput. Syst. Sci. Eng.* **2022**, *42*, 837–849. [CrossRef]
13. Bosquet, B.; Mucientes, M.; Brea, V.M. STDnet: Exploiting high resolution feature maps for small object detection. *Eng. Appl. Artif. Intell.* **2020**, *91*, 103615. [CrossRef]
14. Liu, Z.; Gao, G.; Sun, L.; Fang, Z. HRDNet: High-resolution detection network for small objects. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
15. Yang, C.; Huang, Z.; Wang, N. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13668–13677.
16. Noh, J.; Bae, W.; Lee, W.; Seo, J.; Kim, G. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9725–9734.
17. Ramesh, D.B.; Iytha Sridhar, R.; Upadhyaya, P.; Kamaleswaran, R. Lugsam: A Novel Framework for Integrating Text Prompts to Segment Anything Model (Sam) for Segmentation Tasks of Icu Chest X-Rays. 4 February 2024. Available online: https://ssrn.com/abstract=4676192 (accessed on 30 March 2024).
18. Cen, J.; Zhou, Z.; Fang, J.; Shen, W.; Xie, L.; Jiang, D.; Zhang, X.; Tian, Q. Segment anything in 3d with nerfs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 25971–25990.
19. Réby, K.; Guilhelm, A.; De Luca, L. Semantic Segmentation using Foundation Models for Cultural Heritage: An Experimental Study on Notre-Dame de Paris. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 1689–1697.
20. Li, Y.; Wang, D.; Yuan, C.; Li, H.; Hu, J. Enhancing agricultural image segmentation with an agricultural segment anything model adapter. *Sensors* **2023**, *23*, 7884. [CrossRef] [PubMed]
21. Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H. Grounded sam: Assembling open-world models for diverse visual tasks *arXiv* **2024**, arXiv:2401.14159.
22. Jiao, S.; Wei, Y.; Wang, Y.; Zhao, Y.; Shi, H. Learning mask-aware clip representations for zero-shot segmentation. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 35631–35653.
23. Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; Zhang, L. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 8815–8827
24. Quick Label. Available online: https://github.com/gaoCleo/quick-label (accessed on 10 March 2024).
25. Xu, X.; Geng, Q.; Gao, F.; Xiong, D.; Qiao, H.; Ma, X. Segmentation and counting of wheat spike grains based on deep learning and textural feature. *Plant Methods* **2023**, *19*, 77. [CrossRef] [PubMed]

26. Gao, Y.; Li, Y.; Jiang, R.; Zhan, X.; Lu, H.; Guo, W.; Yang, W.; Ding, Y.; Liu, S. Enhancing green fraction estimation in rice and wheat crops: A self-supervised deep learning semantic segmentation approach. *Plant Phenomics* **2023**, *5*, 0064. [CrossRef] [PubMed]
27. Shen, R.; Zhen, T.; Li, Z. Segmentation of unsound wheat kernels based on improved mask RCNN. *Sensors* **2023**, *23*, 3379. [CrossRef] [PubMed]