remote sensing

MDPI

Article

Cotton Yield Prediction via UAV-Based Cotton Boll Image Segmentation Using YOLO Model and Segment Anything Model (SAM)

Janvita Reddy [1], Haoyu Niu [1,2,*], Jose L. Landivar Scott [3], Mahendra Bhandari [3], Juan A. Landivar [3], Craig W. Bednarz [4] and Nick Duffield [1,2]

[1] Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX 77843, USA
[2] Texas A&M Institute of Data Science, Texas A&M University, College Station, TX 77843, USA
[3] Texas A&M AgriLife Research and Extension Center, Texas A&M University, Corpus Christi, TX 78406, USA
[4] Department of Agricultural Sciences, West Texas A&M University, Canyon, TX 79016, USA
* Correspondence: hniu@tamu.edu

Abstract: Accurate cotton yield prediction is essential for optimizing agricultural practices, improving storage management, and efficiently utilizing resources like fertilizers and water, ultimately benefiting farmers economically. Traditional yield estimation methods, such as field sampling and cotton weighing, are time-consuming and labor intensive. Emerging technologies provide a solution by offering farmers advanced forecasting tools that can significantly enhance production efficiency. In this study, the authors employ segmentation techniques on cotton crops collected using unmanned aerial vehicles (UAVs) to predict yield. The authors apply Segment Anything Model (SAM) for semantic segmentation, combined with You Only Look Once (YOLO) object detection, to enhance the cotton yield prediction model performance. By correlating segmentation outputs with yield data, we implement a linear regression model to predict yield, achieving an R^2 value of 0.913, indicating the model's reliability. This approach offers a robust framework for cotton yield prediction, significantly improving accuracy and supporting more informed decision-making in agriculture.

Keywords: cotton; semantic segmentation; unmanned aerial vehicles; segment anything; yield estimation; object detection

Citation: Reddy, J.; Niu, H.; Scott, J.L.L.; Bhandari, M.; Landivar, J.A.; Bednarz, C.W.; Duffield, N. Cotton Yield Prediction via UAV-Based Cotton Boll Image Segmentation Using YOLO Model and Segment Anything Model (SAM). *Remote Sens.* **2024**, *16*, 4346. https://doi.org/10.3390/rs16234346

Academic Editor: Jianxi Huang

Received: 30 September 2024
Revised: 19 November 2024
Accepted: 19 November 2024
Published: 21 November 2024

1. Introduction

Crop yield prediction is increasingly critical due to concerns about food security and the need to estimate food availability for a growing population [1]. Accurate early predictions are essential for reducing famine risks and improving resource management, which benefits farmers through better decisions on crop insurance, harvest planning, and budgeting [2]. Traditional methods, such as labor-intensive manual surveys and modeling approaches, are often costly, subjective, and inefficient, especially for large-scale operations [3]. Remote sensing methods use vegetation index, such as the normalized difference vegetation index (NDVI), to estimate crop yields [4]. For instance, Anastasiou et al. demonstrated that proximal sensing, particularly green normalized difference vegetation (GNDVI) during veraison, provided stronger correlations with grape yield compared to satellite-based methods [5]. Additionally, Kogan et al. used the advanced very high resolution radiometer (AVHRR)-based vegetation condition index (VCI) to estimate vegetation state and productivity, showing a strong correlation with crop density [6]. However, index-based approaches are limited by their reliance on calibration, manual interpretation, and difficulties in handling environmental variability and high-dimensional data. Unmanned aerial vehicle (UAV) technology has emerged as a valuable tool in ecological research for monitoring vegetation and ecosystems, offering a cost-effective and user-friendly alternative for agricultural remote sensing applications [7–9].

Machine learning and computer vision techniques have been developed for crop yield prediction. For example, Veenadhari et al. developed a classifier using the decision tree approach that helps to understand how different climatic factors influence the results [10]. Ramesh et al. developed statistical models, such as multiple linear regression and a density-based model, to estimate future year yield, taking into account variables such as snow area, fertilizers, rainfall, and production [11]. Khaki et al. trained deep neural network models that incorporate genotype, weather, and soil conditions to predict and check yield [12]. In addition, feature selection was performed to optimize the input space, without significant reduction in accuracy. Aggarwal et al. applied histogram equalization and k-means clustering to separate the crop from the background in images, increasing accuracy and computational efficiency [13]. You et al. effectively integrated spatial and temporal features by combining long short-term memory (LSTM) networks with Gaussian processes for accurate yield prediction of soybean crop [14]. Computer vision tools are powerful for automating the inspection tasks in agriculture. Image segmentation is a crucial step in computer vision tasks where the goal is to partition an image into meaningful segments, typically by classifying each pixel into a predefined category. In the context of agriculture, segmentation plays an essential role in extracting valuable information from images of crops. For example, Wang et al. developed an automated system for apple yield estimation by capturing the nighttime images of trees and segmented the apples based on color cues [15]. Morphological operations were applied to locate and count the apples. In [16], Sarkate et al. performed color segmentation using thresholding and histogram analysis in the hue, saturation, and value (HSV) color space to detect gerbera flowers.

Deep learning-based segmentation has gained significant prominence due to its robustness and adaptability. Unlike traditional methods, which are often sensitive to noise and require carefully tuned thresholding values, deep learning models excel in handling variability within agricultural images [17,18]. For instance, in [19], Palacios et al. focused on predicting the number of berries and the overall yield of grapevines by employing a SegNet-based convolutional neural network (CNN) to detect berries and canopy features in grapevine images, which were then used to train support vector regression (SVR) models for predicting berry count and yield. This approach allowed accurate yield forecasting up to 60 days before harvest. In [10], Veenadhari et al. created an ensemble framework based on the bagging strategy and the UNet network. It utilizes both RGB and HSV color spaces to improve the segmentation of maize crop. In [20], Yu et al. modified UNet by reducing the downsampling rate, and adding attention as well as feature extract blocks to help model distinguish oranges from complex backgrounds in orchard environments. CNNs have become highly significant in segmentation tasks due to their impressive accuracy and effectiveness. However, their ability to capture information is inherently limited to local regions [21], and they struggle to model long-range dependencies across an image. This limitation arises from the localized nature of convolutional operations and the fixed receptive field of convolutional layers. Following the introduction of the self-attention mechanism in [22], there has been a significant surge in the application of Transformer models. The integration of CNNs and Transformers is particularly promising for complex vision tasks that require both fine-grained feature extraction and an understanding of broader spatial relationships. For instance, in [23], Chen et al. proposed TransUnet, which used transformer to encode image tokens from a CNN feature map to extract global contexts, while the decoder upscales these encoded features and merges them with low-level feature maps to attain precise location. In [24], Silva et al. leveraged a hybrid approach by adding a resnet block after the transformer encoder to combine the advantages of both CNN and vision transformer (ViT) for soybean weed detection. In [21], Coro et al. employed an encoder CNN combined with a spatial and channel reconstruction unit to preserve essential spatial information, while developing a decoder transformer with multiple attention mechanisms to focus on the local features of crops.

More recently, the Segment Anything Model (SAM) was proposed, and gained a lot of attention because of its capability to accurately segment the images based on user prompts [25]. This foundational model was trained on a large dataset of 1 million im-

ages and over 1 billion masks, which makes it highly adaptable and able to transfer its knowledge to any new distribution of images. For instance, in [26], Zhang et al. evaluated the performance between SAM automatic mode and SAM box prompt for segmenting clinical radiotherapy images. In [27], Zhang et al. developed SAMed for medical image segmentation, which infused a low-rank adaptation fine-tuning strategy, also incorporating a warm-up strategy and the Adam optimizer for better convergence and lower loss. Though SAM has excellent segmentation capabilities, it is seldom used to segment on agriculture-specific domains. To address this issue, researchers are working on fine-tuning millions of parameter-trained models. In [28], Li et al. incorporated SAM adaptors between the decoder layers, keeping all other parameters constant, to better suit the segmentation task for agricultural images.

Despite the strong zero-shot segmentation capabilities of SAM, it struggles with multiple objects and domain-specific images, such as those from agricultural settings. To overcome these limitations, in this article, the authors explore the innovative use of vision foundational models for cotton yield prediction, which leverages trained prompts for segmenting various objects, and present a comparative analysis showcasing its superior performance against other models on cotton yield prediction. Cotton plays a vital role in the global economy and is one of the most widely used natural fibers worldwide. Its versatility makes it essential in various industries, from textiles to food production. In the United States (US), cotton is a major cash crop that contributes significantly to the agricultural sector. The United States is one of the largest producers and exporters of cotton globally, accounting for about 35% of global cotton production [29]. Among US states, Texas is the largest producer, responsible for approximately 40% of the nation's cotton production in recent years [30]. This makes cotton essential not only for the agricultural economy but also for various industries that rely on its diverse applications.

The objective of this study is to develop and evaluate deep learning models, specifically YOLO and SAM, for cotton boll image segmentation and yield prediction using UAV RGB images. The major contributions are (1) the application of YOLO for high-accuracy cotton boll detection; (2) the integration of the SAM model for semantic segmentation in cotton yield prediction; and (3) the provision of the model framework, code, and methodology to support future research and applications in agriculture. The rest of the manuscript is organized as follows: Section 2 outlines in detail the materials and methods used to predict cotton yield using the YOLO and SAM models. Section 3 delves into a comprehensive analysis and discussion of cotton boll image segmentation achieved with these models, highlighting the relationship between yield and segmentation results. Finally, Section 4 concludes the study, summarizing the key findings and their implications for agricultural research.

2. Materials and Methods

2.1. Study Area

The field experiments were conducted in Lubbock, Texas (33.59°N, 101.90°W) on a 2.2-acre cotton field (Figure 1). Cotton was harvested mechanically from each individual row, generating a dataset comprising 96 rows of yield data. Each row extended 200 feet, containing about 150 cotton plants, with a 40-inch distance between rows [8].

Figure 1. The cotton field on 9 November 2022.

2.2. UAV Image Acquisition and Processing

High-resolution aerial images of the cotton field were acquired on 9 November 2022 using a DJI Phantom 4 Pro (Shenzhen DJI Sciences and Technologies Ltd., Shenzhen, China) equipped with a 1 inch complementary metal oxide semiconductor sensor (CMOS). The flight was carried out at an altitude of 25 m above ground level (AGL) and lasted approximately 18 min. The UAV flight path was planned with an 80% side and front overlap, resulting in a total of 406 high-resolution aerial images.

Data processing was performed with Agisoft Metashape to generate an orthomosaic. The image processing workflow followed a structured pipeline, starting with the alignment of the photos by identifying key points. Ground Control Points (GCPs) were incorporated to improve spatial accuracy, followed by optimizing camera parameters to refine alignment. A dense point cloud was then created, which was further utilized to build a Digital Elevation Model (DEM). Finally, the orthomosaic was generated using the DEM and aligned images. A visual representation of this workflow is presented in Figure 2, which outlines the step-by-step process from image acquisition to the generation of final data products.

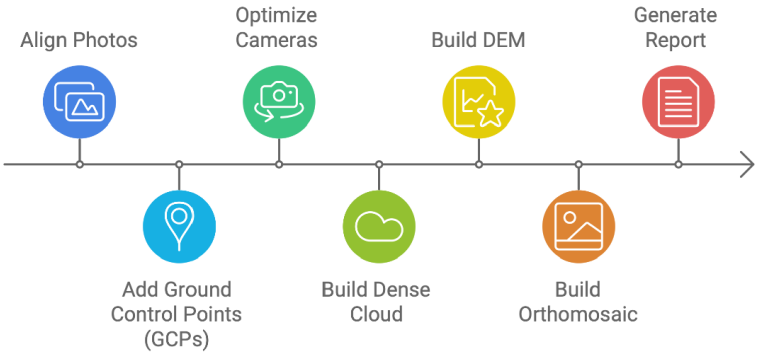

Figure 2. The step-by-step workflow from image acquisition to the generation of the orthomosaic image.

2.3. Image Annotation

The authors used the AnyLabeling software (Version: 0.4.15) to annotate images, leveraging SAM for automatic segmentation. The process involved running the encoder only once for each image, followed by running the decoder based on input prompts such as points or boxes to generate output masks. A post-processing step was automatically implemented to identify contours and created shapes like polygons or rectangles for labeling. AnyLabeling offers three versions of SAM. The original SAM, trained on 11 million images and 1 billion segmentation masks, excels at segmenting objects without prior training knowledge, making it ideal for autolabeling even with new objects. SAM 2, Meta's latest advancement, offers enhanced visual segmentation for both images and videos [31]. MobileSAM is a lightweight variant designed for mobile applications [32]. For our image annotation, we used the original SAM model, annotating around 1800 images. A demonstration of the original images and labeled masks is shown in Figure 3.

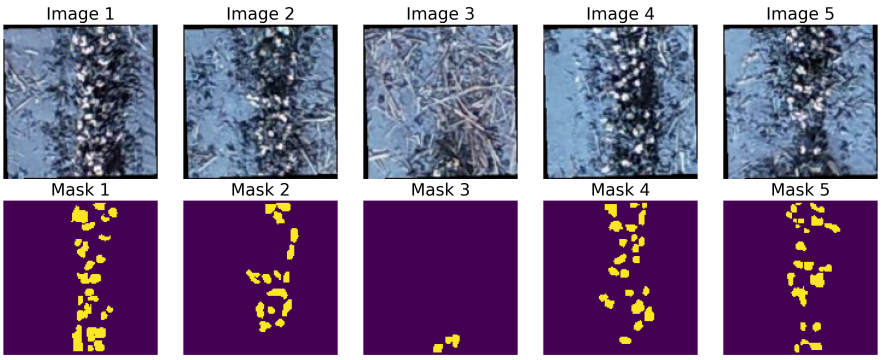

Figure 3. A demonstration of the original and mask images.

2.4. The Segment Anything and YOLO Models

SAM is a prompt-based model that, similar to language models, generates a valid segmentation mask from ambiguous prompts like points, text, or rough boxes. It comprises an image encoder, a pretrained masked autoencoder vision transformer that processes the image input, and a prompt encoder that handles sparse prompts by embedding them with learned embeddings and positional encodings. Text embeddings are generated using the contrastive language-image pretraining (CLIP) encoder, while dense prompts (masks) are processed through convolutional layers [33]. The decoder, inspired by Transformer models, uses self-attention within prompts and cross-attention between prompts and image features to update the embeddings. After two processing blocks, the image features are upscaled, and a neural network maps the output token to a classifier, which predicts the foreground mask, resulting in the final segmentation mask.

The YOLO v7 and YOLO v8 object detection algorithms were mainly explored for generating bounding boxes. The YOLO models have continuously evolved to address the limitations and disadvantages of their previous versions. Unlike traditional models such as RCNN and Fast RCNN [34], which use two separate outputs for object detection, one for classification and the other for bounding-box regression, the YOLO models perform both tasks in a single pass. This streamlined approach allows YOLO to predict class labels and bounding box coordinates simultaneously, enhancing both efficiency and performance [35]. The architecture of YOLO v7 is structured into three main components: the backbone, neck, and head [36]. The backbone is crucial for the extraction of features, playing a significant role in both the training of the model and its overall efficiency. It is responsible for capturing and representing detailed features from the input data. The neck functions as the feature aggregator, combining low-level spatial information with high-level semantic information to build rich feature maps at various levels. Finally, the head utilizes these aggregated features to perform final object detection, generating precise bounding boxes and class predictions. To enhance feature representation, YOLO v7 employs the efficient layer aggregation network (ELAN). Unlike traditional models that use stack convolutional layers, which limit detection capabilities for objects of various sizes, ELAN aggregates features from multiple layers. This approach allows YOLO v7 to combine information from different stages of processing, capturing features at multiple scales simultaneously. ELAN further extends this by integrating multiple paths for feature aggregation, improving the model's ability to handle a broad range of features without significantly increasing computational costs. YOLO v8 is the current superior object detection model built upon its previous versions to boost the performance and flexibility. Along with object detection, pose estimation, tracking and classification, this architecture also supports segmentation. Unlike earlier versions that relied on anchor boxes for detecting objects, YOLO v8 utilizes an anchor-free approach, which does not use predefined boxes [37]. It directly predicts the location and size of the objects. This reduces the computational complexity and makes the model simpler and faster, with fewer hyperparameters to tune. The head of the network is "decoupled"; it separates the processing for the objectness score, classification, and regression, which lets the model specialize in its specific task.

In this article, we propose an innovative two-stage approach for image segmentation that integrated object detection with advanced segmentation technique. In the first stage, we employed YOLO v7 and YOLO v8 as the object detector, which processed the input images to predict bounding boxes around the region of interest (Figure 4). These bounding boxes were then passed as input to the prompt encoder, which performed prompt encoding by embedding the box coordinates along with positional encodings. More specifically, from the AnyLabeling software, we extracted the bounding box coordinates for the bolls in the cotton images. These bounding boxes were then converted to the YOLO labeling format using Roboflow, which transformed the coordinates into the required format: (class index, normalized x center, normalized y center, normalized width, normalized height). Since all of the images contained only one class, cotton bolls, the class index was consistently set to 0. Out of the 1800 images, we used 1500 for training, 100 for validation, and the remaining 200 for testing the model. To further enhance the performance of the model,

we resized the image resolution from 150 × 150 to 640 × 640 (on which originally YOLO models were trained) pixels to improve object detection for the small cotton bolls scattered across the field.

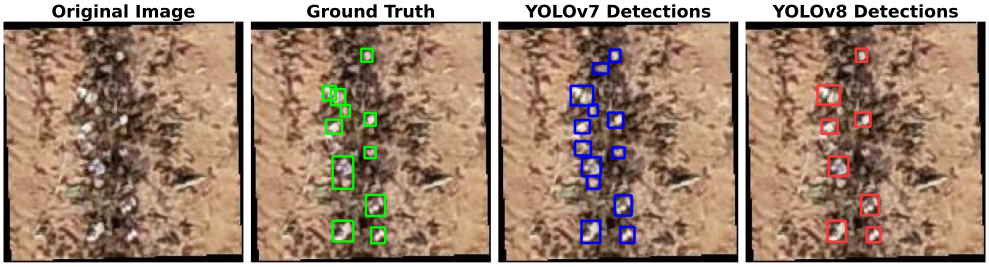

Figure 4. Comparison of cotton boll detection between YOLO v7 and YOLO v8 models.

In the second stage, the SAM encoded the entire image using a pretrained vision transformer. The bounding boxes, now encoded by the prompt encoder, were also fed into the SAM decoder (Figure 5). The SAM decoder performed a series of operations involving both self-attention within the prompts and cross-attention between the image features and the encoded prompts. This bidirectional attention mechanism enabled the model to refine and update the embeddings, ensuring that the segmentation was accurate and aligned with the detected objects. This information was processed to generate the final output, a segmented mask that precisely delineated the objects within the image. By combining YOLO v7 or v8 for object detection with SAM's powerful segmentation capabilities, our approach effectively handled images with complex scenarios, resulting in a robust and efficient solution for image segmentation tasks.

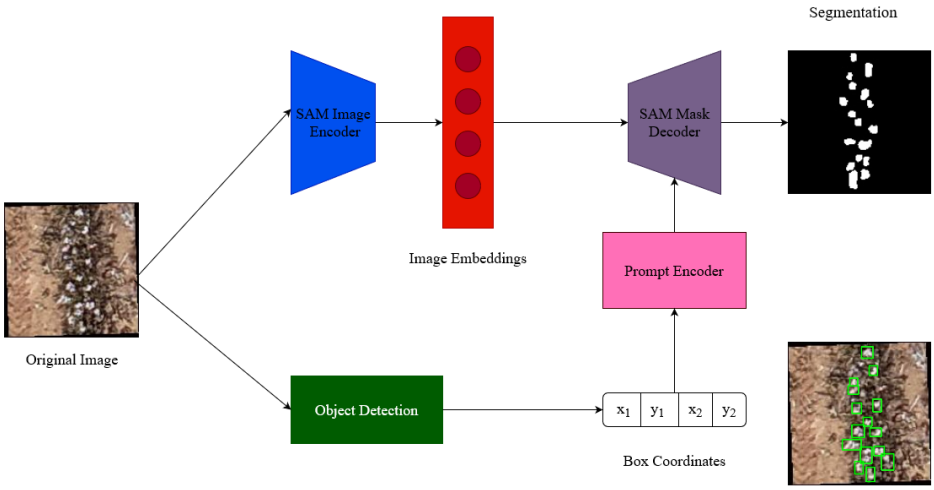

Figure 5. The proposed YOLO + SAM model architecture.

2.5. Model Evaluation Metrics

To evaluate the model's performance in detecting objects, we use the following metrics. Precision measures the proportion of objects detected by the model that are actually relevant (true positives). It is calculated as the ratio of true positives (TP) to the sum of true positives and false positives (FP). Precision reflects how accurate the detections are by indicating the percentage of correct detections (Equation (1)). Recall measures how well the model detects relevant objects in the dataset. It is calculated as the ratio of true positives to the sum of true positives and false negatives (FN). Recall indicates the model's ability to identify all relevant objects (Equation (2)). F1-Score finds the most optimal confidence score threshold where precision and recall give the highest F1 score. The F1 score calculates the balance between precision and recall. If the F1 score is high, precision and recall are high, and vice versa (Equation (3)):

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{2}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{3}$$

To evaluate the performance of our segmented images, we use the Intersection over Union (IoU) score, also known as the Jaccard coefficient (Equation (4)). The IoU score measures the overlap between the ground truth mask A and the predicted mask B, providing a quantitative measure of the precision of the segmentation of objects in the region of interest. The area of intersection is the area where the predicted mask and the ground truth mask overlap, and the area of union is the total area covered by both the predicted mask and the ground truth mask combined. For Mean Average Precision (mAP), the IoU threshold is set to 0.5, which means that a predicted object is considered a true positive if the overlap between the predicted bounding box and the ground truth bounding box is at least 50%. It indicates how competent the model is in localizing the objects:

$$\text{Jaccard}(A, B) = \frac{\text{Area of Intersection } (A \cap B)}{\text{Area of Union } (A \cup B)}. \tag{4}$$

3. Results and Discussions

3.1. The SAM and YOLO Model Performance

To achieve an optimal balance between precision and recall, we systematically evaluate the the F1 score on various decision thresholds. By plotting the F1 score as a function of the confidence threshold, we identify the threshold that maximizes the F1 score for each model on the validation dataset. For YOLO v7, the optimal confidence threshold is determined to be 0.3, while for YOLO v8, a threshold of 0.25 yields the highest F1 score. This analysis reveals that the maximum achievable F1 score in both models is 0.83. These thresholds ensure that the models achieve an optimal trade-off between precision and recall, enhancing their ability to accurately identify relevant objects while minimizing false positives and false negatives. The output of each object in the image includes the class number and the normalized coordinates of its bounding box, represented as $(x_{\text{center}}, y_{\text{center}}, w, h)$, where x_{center} and y_{center} are the normalized coordinates of the center of the bounding box, and w and h are the normalized width and height. To convert these normalized coordinates to absolute pixel coordinates, we use the following equations:

$$\begin{aligned} x_{\min} &= \left(x_{\text{center}} - \frac{w}{2}\right) \times W, \qquad y_{\min} = \left(y_{\text{center}} - \frac{h}{2}\right) \times H; \\ x_{\max} &= \left(x_{\text{center}} + \frac{w}{2}\right) \times W, \qquad y_{\max} = \left(y_{\text{center}} + \frac{h}{2}\right) \times H; \end{aligned} \tag{5}$$

where $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ represent the coordinates of the bottom-left and top-right corners of the bounding box in pixel coordinates, respectively. The width W and height H refer to the dimensions of the image in pixels. These absolute coordinates are then provided to the SAM as prompts for segmenting the regions around the bounding box. The SAM model uses these coordinates to accurately identify and segment objects of interest within the specified bounding box, ensuring precise and focused segmentation. From the comparative analysis presented in Table 1, it is evident that YOLO v7 outperforms YOLO v8 in several key metrics. YOLO v7 shows higher precision, recall, F1 score, and mAP@50 compared to YOLO v8. Furthermore, YOLO v7, when combined with SAM, achieves a higher IoU score of 0.683. This could be justified because there are more true positives detected in YOLO v7 compared to YOLO v8. These results indicate that YOLO v7 is more effective in accurately detecting and localizing objects, making it a more robust

choice for object detection tasks. Also, from Figure 6, we observe that YOLO v7 not only detects all objects in the image but also identifies objects that are missed during labeling, further validating the generalization capabilities of our algorithm. Furthermore, YOLO v7 excels in locating smaller objects, which is a significant advantage over YOLO v8. YOLO v8 prioritizes certain optimizations such as faster inference or reduced model complexity, which leads to a slight compromise in detection accuracy, especially in specific tasks like detecting small objects. On the other hand, YOLO v7 incorporates architectural enhancements and reparameterization techniques that are particularly effective in improving precision, recall, and localization accuracy. This makes YOLO v7 better suited for cotton boll detection, where high detection accuracy and robust performance are more critical than other considerations like speed or model size. Therefore, while YOLO v8 brings advances, it might not always outperform YOLO v7 in terms of raw detection performance.

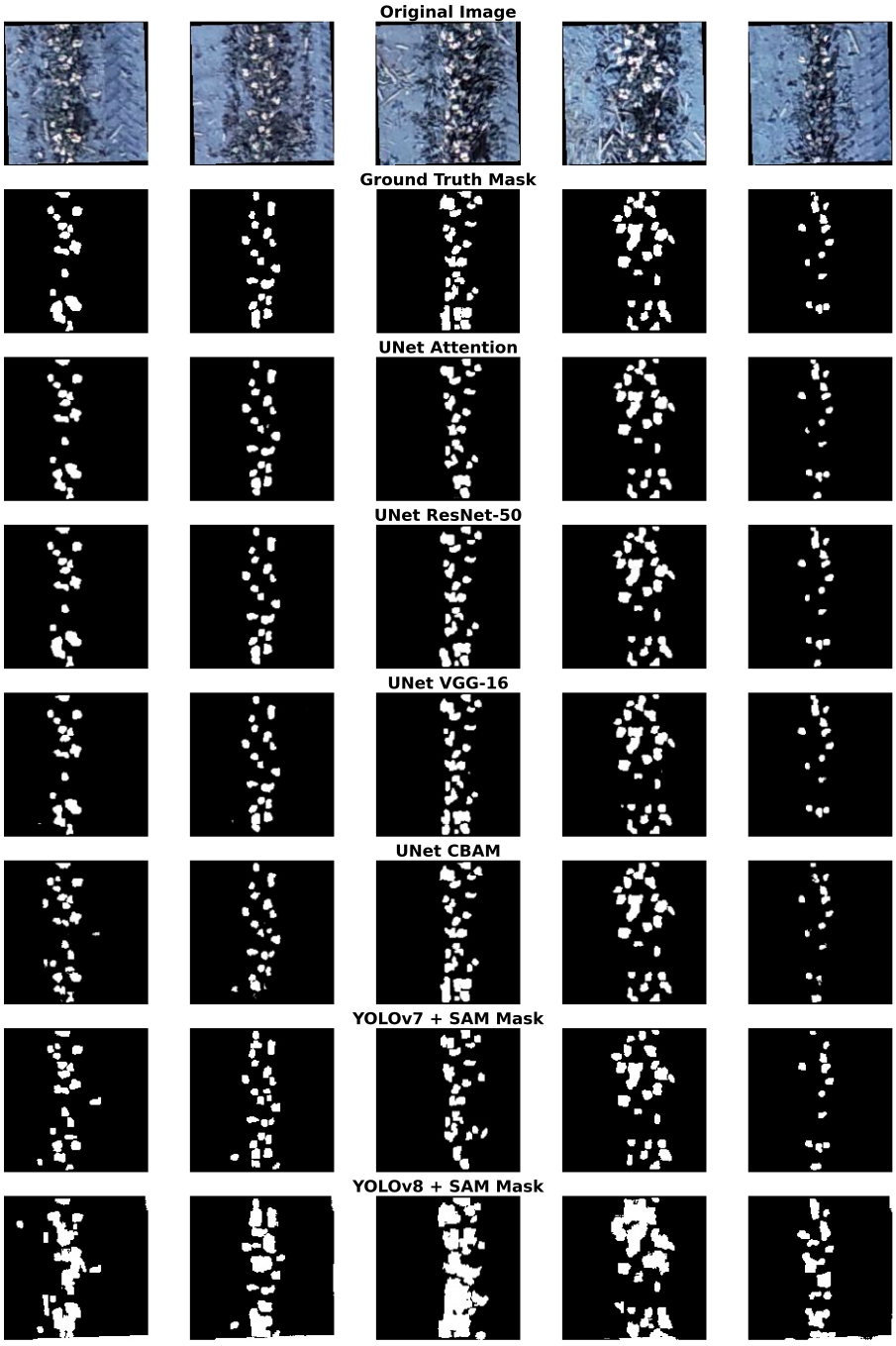

Figure 6. Comparison of cotton boll segmentation between different models.

Table 1. Comparison of YOLO v7 and YOLO v8 metrics.

Models	Precision	Recall	F1-Score	$mAP_{0.5}$	IoU
YOLO v7 + SAM	0.821	0.836	0.828	0.857	0.685
YOLO v8 + SAM	0.814	0.791	0.802	0.833	0.683

In addition to its superior object detection capabilities, the integration of YOLO v7 with SAM for segmentation tasks demonstrates significant improvements in both efficiency and accuracy compared to conventional segmentation models. Traditional segmentation models require a large amount of labeled data for training and struggle with small or overlapping objects. In contrast, our approach leverages the precise bounding boxes generated by YOLO v7 to provide focused and accurate segmentation prompts to SAM. This results in a more computationally efficient pipeline, as the model narrows down the regions of interest before applying segmentation. Furthermore, the high IoU score of 0.685 indicates that the YOLO v7 + SAM combination is not only faster but also more effective in identifying object boundaries, particularly for smaller objects that traditional models often miss. This combination of speed and accuracy makes our model a suitable choice for real-time applications requiring both object detection and segmentation, offering a substantial advantage over conventional segmentation models. We experimented with well-known U-Net architecture models for segmentation. U-Net with attention mechanisms leverages the encoder–decoder structure to focus on specific regions. Variants like U-Net with ResNet 50 and VGG 16 utilize models pretrained on the ImageNet dataset as their encoders, benefiting from feature extraction capabilities. Additionally, U-Net with a Convolutional Block Attention Module (CBAM) introduces channel and spatial attention in the encoder, further refining the focus on relevant features. Table 2 presents the efficiency metrics of each model configuration.

Table 2. Comparison of multiple segmentation models.

Models	IoU Score	Inference Time (s)
U-Net Attention	0.697	0.0087
U-Net ResNet 50	0.696	0.033
U-Net VGG 16	0.685	0.0128
U-Net CBAM	0.680	0.013
Yolov7 + SAM	0.685	0.526
Yolov8 + SAM	0.683	0.554

3.2. Evaluation of Cotton Yield at Row Level

We apply image segmentation using the proposed method to segment cotton bolls from field images. The segmentation process allows us to isolate cotton bolls as white pixels in the images. We then count the number of white pixels in each image, which represent the presence of cotton. To estimate the cotton yield for each row, we aggregate the white pixel counts across all images belonging to that specific row. This approach allows us to derive an estimated yield for the entire row on the basis of the segmented images. Next, we aim to quantify the relationship between the white pixel count and the actual cotton yield. A linear regression model is applied, where the white pixel count is used as the independent variable and the actual yield as the target variable. We use data from 58 rows to construct the linear regression model and test it on data from 38 rows. In Figure 7, we plot the pixel count for each row on the x axis and the actual yield on the y axis. We observe that the model successfully captures the correlation between these two values, achieving a high R^2 value of 0.913 with YOLOv8 + SAM, indicating that the model explains a large proportion of the variance in cotton yield. The model has a low mean absolute error (MAE) of 3.872 lbs/row, showing that the predictions are accurate.

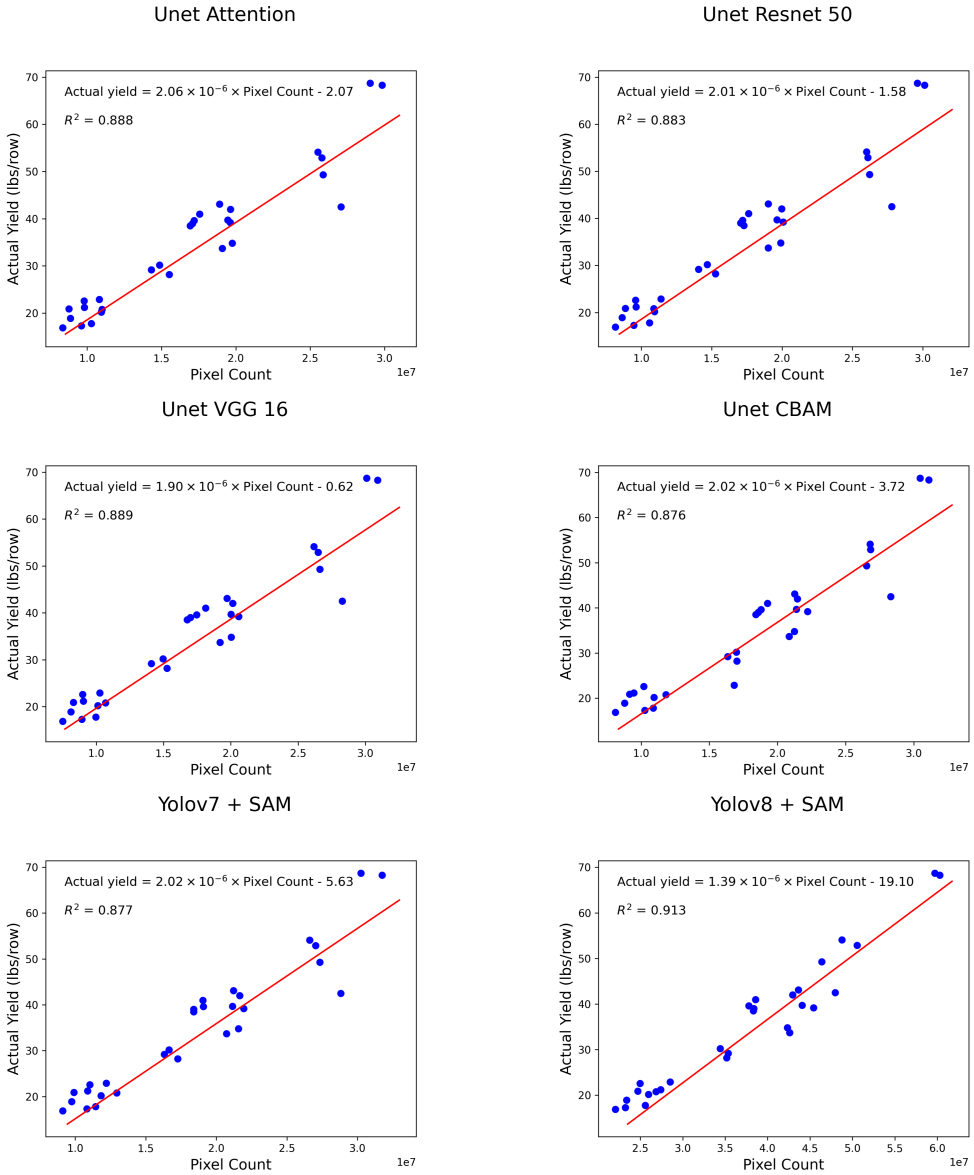

Figure 7. Linear relationship between pixel count and actual yield across different models. The YOLOv8 + SAM has the best yield prediction performance, with a R^2 of 0.913.

However, it is important to note that a higher IoU score, which typically indicates better segmentation quality, does not necessarily translate to higher accuracy in yield prediction (Figure 7). While a higher IoU reflects a better match between the predicted and ground truth segmentation, it does not always correlate with the overall cotton yield since other factors, such as boll distribution, growth stage, and environmental conditions, can influence yield independently of segmentation accuracy. Thus, a high IoU score might not fully capture the complexities of yield estimation. In contrast, our linear regression model focuses on the aggregated pixel counts, which are more directly related to the overall yield, demonstrating that effective yield prediction depends not only on accurate segmentation but also on how well the segmentation relates to actual cotton production. Future studies may explore additional features or hybrid models that consider both segmentation accuracy and environmental variables to improve yield predictions.

4. Conclusions

In this study, the authors proposed a two-stage approach for image segmentation that integrates object detection with an advanced segmentation technique to predict cotton

yield. A comparison between two recent state-of-the-art object detection models, YOLO v7 and YOLO v8, was conducted. YOLO v7 was observed to perform better in all metrics, and this resulted in a high IoU score of 0.685. Using YOLO v8 and SAM in conjunction with UAV imagery, we were able to achieve a high correlation between the segmented output and actual yield data, reflected in an R^2 value of 0.913. This model provides a robust and reliable method for predicting cotton yields, significantly reducing the need for traditional labor-intensive methods. The proposed approach not only improves prediction accuracy but also offers a practical tool to optimize agricultural practices, resource allocation, and decision-making processes. It paves the way for exploring the potential of vision foundational models in agricultural applications. Additionally, the combination of object detection with segmentation in our approach aligns with the growing interest in zero-shot or few-shot segmentation, which enables models to adapt to new crop types with minimal labeled data. Future work could explore the application of this framework to other types of crops and further refine the model by incorporating additional variables, such as soil health and climatic conditions, to improve its predictive power and generalizability in diverse agricultural environments. In this study, we focused on a specific dataset collected in the Lubbock, Texas area, with parameters optimized based on the characteristics of this location, cotton variety, and growth stage. While this approach was effective within the defined scope, we agree that further research is needed to explore how segmentation parameters may vary across different planting sites, cotton species, growth stages, and imaging techniques. As part of future work, we intend to investigate how our segmentation model generalizes under varied conditions and to test parameter adjustments that may enhance its adaptability.

5. Research Reproducibility

We agree that generalizing to other regions and environmental conditions would be valuable, and we have noted this as a future research direction. We believe that the model's framework can be adapted to different regions by local researchers. To support this, we have provided a detailed description of our methodology and made the model code available, facilitating adaptation and validation in diverse contexts. All the research results in this article can be reproduced. The code is available in the author's Github: https://github.com/hniu-tamu/Cotton_yield_prediciton_with_YOLO_and_SAM, accessed on 10 November 2024.

Author Contributions: Conceptualization, H.N. and J.R.; methodology, H.N. and J.R.; software, H.N. and J.R.; validation, H.N. and J.R.; formal analysis, H.N. and J.R.; investigation, J.R.; resources, M.B., J.A.L., C.W.B. and N.D.; data curation, J.R. and J.L.L.S.; writing—original draft preparation, J.R., J.L.L.S. and H.N.; writing—review and editing, J.R., J.L.L.S. and H.N.; visualization, J.R., J.L.L.S. and H.N.; supervision, H.N. and N.D.; project administration, C.W.B., M.B., J.A.L. and N.D.; funding acquisition, C.W.B., M.B., J.A.L. and N.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Texas state allocated funds for the Water Exceptional Item through Texas A&M AgriLife Research facilitated by the Texas Water Resources Institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors sincerely appreciate Murilo Maeda's invaluable assistance as the project lead responsible for collecting all the data used in this experiment during 2022.

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

Abbreviations

The following abbreviations are used in this manuscript:

AGL	Above Ground Level
AVHRR	Advanced Very High Resolution Radiometer
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Convolutional Neural Network
CLIP	Contrastive Language-Image Pretraining
DEM	Digital Elevation Model
ELAN	Efficient Layer Aggregation Network
FN	False Negative
FP	False Positive
GCPs	Ground Control Points
GNDVI	Green Normalized Difference Vegetation Index
HSV	Hue, Saturation, and Value
IoU	Intersection over Union
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
mAP	Mean Average Precision
NDVI	Normalized Difference Vegetation Index
RGB	Red, Green and Blue
SAM	Segment Anything Model
SVR	Support Vector Regression
TP	True Positive
UAV	Unmanned Aerial Vehicle
US	United States
VCI	Vegetation Condition Index
ViT	Vision Transformer
YOLO	You Only Look Once

References

1. Muruganantham, P.; Wibowo, S.; Grandhi, S.; Samrat, N.H.; Islam, N. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* **2022**, *14*, 1990. [CrossRef]
2. Zhang, M.; Feng, A.; Zhou, J.; Lü, X. Cotton yield prediction using remote visual and spectral images captured by UAV system. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 91–98.
3. Khaki, S.; Pham, H.; Wang, L. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Sci. Rep.* **2021**, *11*, 11132. [CrossRef] [PubMed]
4. Quarmby, N.; Milnes, M.; Hindle, T.; Silleos, N. The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. *Int. J. Remote Sens.* **1993**, *14*, 199–210. [CrossRef]
5. Anastasiou, E.; Balafoutis, A.; Darra, N.; Psiroukis, V.; Biniari, A.; Xanthopoulos, G.; Fountas, S. Satellite and proximal sensing to estimate the yield and quality of table grapes. *Agriculture* **2018**, *8*, 94. [CrossRef]
6. Kogan, F.; Gitelson, A.; Zakarin, E.; Spivak, L.; Lebed, L. AVHRR-based spectral vegetation index for quantitative assessment of vegetation state and productivity. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 899–906. [CrossRef]
7. Ali, A.M.; Abouelghar, M.; Belal, A.; Saleh, N.; Yones, M.; Selim, A.I.; Amin, M.E.; Elwesemy, A.; Kucher, D.E.; Maginan, S.; et al. Crop yield prediction using multi sensors remote sensing. *Egypt. J. Remote Sens. Space Sci.* **2022**, *25*, 711–716.
8. Niu, H.; Peddagudreddygari, J.R.; Bhandari, M.; Landivar, J.A.; Bednarz, C.W.; Duffield, N. In-season cotton yield prediction with scale-aware convolutional neural network models and unmanned aerial vehicle RGB imagery. *Sensors* **2024**, *24*, 2432. [CrossRef]
9. Niu, H.; Chen, Y. *Smart Big Data in Digital Agriculture Applications*; Springer: Berlin/Heidelberg, Germany, 2024.
10. Veenadhari, S.; Mishra, B.; Singh, C. Soybean productivity modelling using decision tree algorithms. *Int. J. Comput. Appl.* **2011**, *27*, 11–15. [CrossRef]
11. Ramesh, D.; Vardhan, B.V. Analysis of crop yield prediction using data mining techniques. *Int. J. Res. Eng. Technol.* **2015**, *4*, 47–473.
12. Khaki, S.; Wang, L. Crop yield prediction using deep neural networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef] [PubMed]
13. Aggarwal, A.K.; Jaidka, P. Segmentation of crop images for crop yield prediction. *Int. J. Biol. Biomed.* **2022**, *7*, 40–44.
14. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian process for crop yield prediction based on remote sensing data. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31.

15. Wang, Q.; Nuske, S.; Bergerman, M.; Singh, S. Design of crop yield estimation system for apple orchards using computer vision. In Proceedings of the 2012 Dallas, Dallas TX, USA, 29 July–1 August 2012; American Society of Agricultural and Biological Engineers: St. Joseph, MI, USA, 2012; p. 1.
16. Sarkate, R.S.; Kalyankar, N.; Khanale, P. Application of computer vision and color image segmentation for yield prediction precision. In Proceedings of the 2013 International Conference on Information Systems and Computer Networks, Mathura, India, 9–10 March 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 9–13.
17. Maji, A.K.; Marwaha, S.; Kumar, S.; Arora, A.; Chinnusamy, V.; Islam, S. SlypNet: Spikelet-based yield prediction of wheat using advanced plant phenotyping and computer vision techniques. *Front. Plant Sci.* **2022**, *13*, 889853. [CrossRef] [PubMed]
18. Peng, H.; Xue, C.; Shao, Y.; Chen, K.; Xiong, J.; Xie, Z.; Zhang, L. Semantic segmentation of litchi branches using DeepLabV3+ model. *IEEE Access* **2020**, *8*, 164546–164555. [CrossRef]
19. Palacios, F.; Diago, M.P.; Melo-Pinto, P.; Tardaguila, J. Early yield prediction in different grapevine varieties using computer vision and machine learning. *Precis. Agric.* **2023**, *24*, 407–435. [CrossRef]
20. Yu, C.; Lin, D.; He, C. ASE-UNet: An orange fruit segmentation model in an agricultural environment based on deep learning. *Opt. Mem. Neural Netw.* **2023**, *32*, 247–257.
21. Corò, F.; D'Angelo, G.; Velaj, Y.; et al. Recommending links to maximize the influence in social networks. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019), Macao, China, 10–16 August 2019; AAAI Press: Washington, DC, USA, 2019; Volume 4, pp. 2195–2201.
22. Vaswani, A. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
23. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
24. Silva, L.; Drews, P.; de Bem, R. Soybean weeds segmentation using VT-Net: A convolutional-transformer model. In Proceedings of the 2023 36th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Rio Grande, Brazil, 6–9 November 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 127–132.
25. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4015–4026.
26. Zhang, L.; Liu, Z.; Zhang, L.; Wu, Z.; Yu, X.; Holmes, J.; Feng, H.; Dai, H.; Li, X.; Li, Q.; et al. Segment anything model (SAM) for radiation oncology. *arXiv* **2023**, arXiv:2306.11730.
27. Zhang, K.; Liu, D. Customized segment anything model for medical image segmentation. *arXiv* **2023**, arXiv:2304.13785.
28. 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]
29. Ridley, W.; Devadoss, S. Competition and trade policy in the world cotton market: Implications for US cotton exports. *Am. J. Agric. Econ.* **2023**, *105*, 1365–1387. [CrossRef]
30. Adhikari, P.; Ale, S.; Bordovsky, J.P.; Thorp, K.R.; Modala, N.R.; Rajan, N.; Barnes, E.M. Simulating future climate change impacts on seed cotton yield in the Texas High Plains using the CSM-CROPGRO-Cotton model. *Agric. Water Manag.* **2016**, *164*, 317–330. [CrossRef]
31. Ravi, N.; Gabeur, V.; Hu, Y.T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment anything in images and videos. *arXiv* **2024**, arXiv:2408.00714.
32. Zhang, C.; Han, D.; Qiao, Y.; Kim, J.U.; Bae, S.H.; Lee, S.; Hong, C.S. Faster segment anything: Towards lightweight SAM for mobile applications. *arXiv* **2023**, arXiv:2306.14289.
33. 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.
34. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
35. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
36. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
37. Varghese, R.; Sambath, M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.