agriculture

MDPI

Article

YOLOv8MS: Algorithm for Solving Difficulties in Multiple Object Tracking of Simulated Corn Combining Feature Fusion Network and Attention Mechanism

Yuliang Gao [1], Zhen Li [2], Bin Li [3] and Lifeng Zhang [1,*]

[1] Graduate School of Engineering, Kyushu Institute of Technology, Kitakyushu 804-0015, Japan
[2] School of Electrical Engineering, Nantong University, Nantong 226021, China
[3] College of Artificial Intelligence, Yangzhou University, Yangzhou 225012, China
* Correspondence: zhang@elcs.kyutech.ac.jp

Abstract: The automatic cultivation of corn has become a significant research focus, with precision equipment operation being a key aspect of smart agriculture's advancement. This work explores the tracking process of corn, simulating the detection and approach phases while addressing three major challenges in multiple object tracking: severe occlusion, dense object presence, and varying viewing angles. To effectively simulate these challenging conditions, a multiple object tracking dataset using simulated corn was created. To enhance accuracy and stability in corn tracking, an optimization algorithm, YOLOv8MS, is proposed based on YOLOv8. Multi-layer Fusion Diffusion Network (MFDN) is proposed for improved detection of objects of varying sizes, and the Separated and Enhancement Attention Module (SEAM) is introduced to tackle occlusion issues. Experimental results show that YOLOv8MS significantly enhances the detection accuracy, tracking accuracy and tracking stability, achieving a mean average precision (mAP) of 89.6% and a multiple object tracking accuracy (MOTA) of 92.5%, which are 1% and 6.1% improvements over the original YOLOv8, respectively. Furthermore, there was an average improvement of 4% in the identity stability indicator of tracking. This work provides essential technical support for precision agriculture in detecting and tracking corn.

Keywords: corn; multiple object track; feature fusion; attention mechanism; YOLOv8

Citation: Gao, Y.; Li, Z.; Li, B.; Zhang, L. YOLOv8MS: Algorithm for Solving Difficulties in Multiple Object Tracking of Simulated Corn Combining Feature Fusion Network and Attention Mechanism. *Agriculture* **2024**, *14*, 907. https://doi.org/10.3390/agriculture14060907

Academic Editor: Marcus Randall

Received: 14 May 2024
Revised: 5 June 2024
Accepted: 7 June 2024
Published: 8 June 2024

1. Introduction

Corn is an important staple food globally [1], and many studies focus on the automatic cultivation of corn [2,3]. Precision agriculture devices, such as robotic arms, can perform fine operations like spot sampling, spraying, and pollination in corn cultivation. These devices offer advantages such as reduced damage, lower losses, and minimal manual intervention. A critical task in this process is the accurate and stable detection and approach of corn, which hinges on multiple object tracking (MOT).

MOT has extensive applications in agriculture [4]. Research has explored tracking crops using drones or unmanned vehicles. For instance, Hu et al. [5] used YOLO-V5 and LettuceTrack to track lettuce in standardized farmland, while Yang et al. [6] employed Centernet+DeepSORT to count cotton by tracking. Villacres et al. [7] utilized various classical algorithms to track and count apples, and Ariza et al. [8] tracked grapes and extracted phenotypes using drone data. Wang et al. [9] used a YOLOv3 network and Kalman filter to count corn seedlings online. However, there is a lack of research on tracking corn fruits. In corn fields, MOT faces challenges such as severe occlusion, dense object presence, and varying viewing angles. This work aims to improve MOT algorithms to achieve better tracking results under these challenging conditions.

MOT is a crucial task in computer vision [10], involving the localization of multiple objects and the maintenance of their identities. The current mainstream method, tracking-by-detection, involves detecting objects in video frames before tracking. During tracking, the algorithm assigns the same identity to bounding boxes detecting the same object. Modern MOT algorithms use advanced detection frameworks to ensure high detection quality and classical tracking algorithms to enhance tracking correlation.

Bewley et al. [11] were pioneers in using convolutional neural networks for object tracking, employing Faster R-CNN as the backbone detection network and combining it with the Kalman filter and Hungarian algorithm for tracking. Milan et al. [12] proposed MOT16 and improved Faster R-CNN for better results. Zhang et al. [13] used an enhanced YOLOv3 to detect and track vehicles, using deep learning frameworks to extract object appearance features and perform nearest neighbor matching, similar to image matching [14]. The tracking algorithm DeepSORT [15] improved robustness by incorporating a recognition algorithm to extract appearance features, enabling a comparison between current and previously stored features. Bytetrack [16] enhances tracking by leveraging the detection box and tracking similarity, retaining high-confidence detection results, and eliminating background noise of low-confidence detections, surpassing DeepSORT in benchmark datasets.

For this work, considering model size and robustness, the YOLOv8 model was chosen as the base model. You Only Look Once (YOLO) is a classical one-stage object detection algorithm known for its rapid detection speed while maintaining precision [17]. To further investigate challenges in MOT of corn, this work created a simulated corn MOT dataset. Simulated corn allows for easier reproduction of challenging conditions compared to real corn fields. Three auxiliary tracking datasets were generated to evaluate model performance, including severe occlusion, dense object presence, and varying viewing angles. Data enhancement techniques, particularly color conversion, were employed to address differences between simulated and real-world corn and background colors.

This work proposed YOLOv8MS, designed to enhance tracking accuracy and stability under challenging conditions. The proposed approach integrates Multi-layer Fusion Diffusion Network (MFDN) and Separated and Enhancement Attention Module (SEAM) [18] to redesign the neck and head networks. The MFDN can improve detection accuracy by accommodating varying corn bounding box sizes. It can fuse context features across multiple scales leveraging the inception mechanism [19] and diffuse its output to different layers. The SEAM module can effectively manage occlusion by strengthening channel connections.

The main contributions of this paper are as follows:

(1) Development of a simulated corn MOT dataset, along with three auxiliary datasets to study specific challenges: severe occlusion, dense object presence, and varying viewing angles. Data enhancement, particularly color conversion, was used to mitigate differences between simulated and real corn.
(2) Proposal of YOLOv8MS, incorporating MFDN in the neck layer and SEAM module in the head layer, to address occlusion and varying object sizes during tracking. Experiments were conducted with various models on complete and auxiliary datasets, evaluating YOLOv8MS's performance in detection and tracking using multiple indicators. Results demonstrate that YOLOv8MS improves not only accuracy but also the stability of tracking, particularly under challenging conditions.

2. Materials and Methods

2.1. Data Collection and Annotation

This work created a MOT dataset using simulated corn, simulating the process of finding and approaching corn using agricultural equipment. Additionally, three auxiliary datasets were constructed to simulate severe occlusion, dense object presence, and varying viewing angles, as shown in Figure 1. These auxiliary datasets help us delve deeper into studying challenging conditions. During data collection, the ZED2I depth camera was used as the data acquisition equipment. A total of 50 sets of raw data were collected, with each

set consisting of videos accompanied by depth maps and camera poses. Each video was recorded at 30 frames per second and averaged about 6 s in length.

For labeling, X-AnyLabeling tools [20] were used. The data were divided as shown in Figure 2. Sixty percent of the video data was annotated in YOLO format and converted into image format by frame. These images were augmented and then divided into training and test sets in a 9:1 ratio and used to train and test the detection model. The remaining 40% of the video data was reserved for testing the algorithm's object tracking capabilities. The tracking video dataset included the three challenging condition auxiliary datasets. The specific composition of the data is shown in Table 1.

The construction details of the three auxiliary datasets are as follows:

(1) Severe Occlusion Auxiliary Dataset:
The corn is first completely shielded by leaves, the video captures the process from complete invisibility to full visibility and then to invisibility again.

(2) Dense Object Presence Auxiliary Dataset:
The simulated corn are placed in a stack, and the video captures the process of approaching them.

(3) Varying Viewing Angles Auxiliary Dataset:
The video captures the process of approaching the simulated corn at different angles. In the video, the relative angles of the corn change significantly as the camera moves.

(A) Severe Occlusion (B) Dense Object Presence (C) Varying Viewing Angles

Figure 1. Samples from auxiliary datasets.

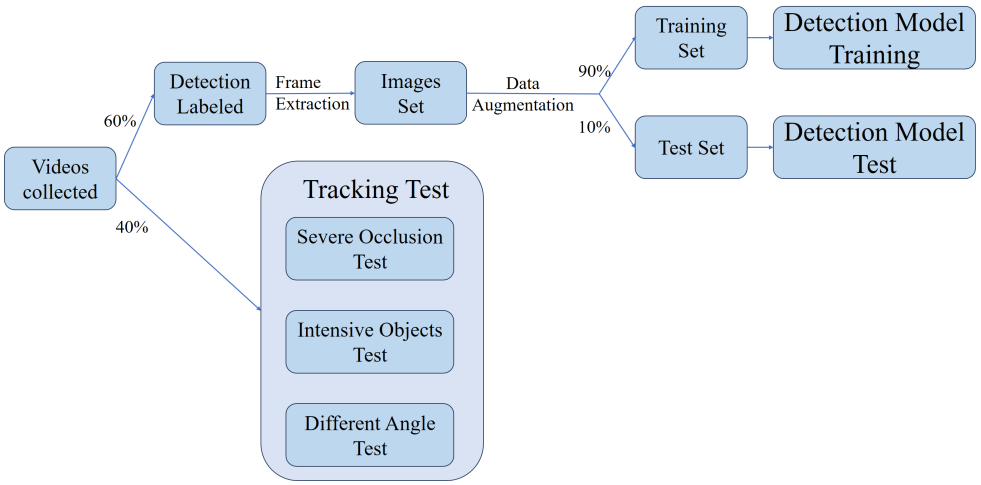

Figure 2. The process of collecting and labeling data for training and testing.

Table 1. This table shows the raw data captions used for training and testing. Instances representing the number of objects used for object detection training and testing. Tracks represent the number of objects used for tracking.

	Detection Training	Detection Test	Tracking Test	Severe Occlusion	Dense Object Presence	Varying Viewing Angles
Resolution	1280 × 720	1280 × 720	1280 × 720	1280 × 720	1280 × 720	1280 × 720
Length (Frames)	4928	492	3328	616	656	656
Instances	18,512	1848	-	-	-	-
Tracks	-	-	66	4	28	12
Application	Train	Detection Test	Test	Auxiliary Test	Auxiliary Test	Auxiliary Test

2.2. Data Augmentation

To improve the generalization of the model and mitigate the limitations of using simulated corn, various data augmentation techniques were applied to the detection training and test. During training, the augmentation is carried out randomly. Before testing, only color conversion is applied on the entire detection test set to expand the test set. As shown in Table 2, the data augmentation methods include mosaic, horizontal flip, scaling (0.75), and color conversion. During data augmentation, color conversion was performed to change the white parts of the original images to green, thereby reducing the difference between the background and the simulated corn and making it more closer to real-world scenarios. Figure 3 provides examples of both augmented and raw images.

Table 2. Table of probabilities for data augmentation.

	Mosaic	Horizontally Flip	Scale (0.75)	Color Conversion
Probability	0.5	0.8	0.5	0.5

Figure 3. Samples of raw and augmented images.

2.3. Experimental Setting

The software and hardware configurations for model training and testing in this work are listed in Table 3. The training epochs are set to 1000 with a batch size of 18. Optimization is performed using the Adam optimizer with a momentum of 0.9. The initial learning rate is set to 0.001, and the learning rate is adjusted every 300 rounds.

Table 3. Software and hardware configuration.

Accessories	Model
CPU	Intel(R) Xeon(R) CPU E5-1650 v4
RAM	64G
Operating system	Ubuntu18.04
GPU	NVIDIA GeForce RTX 1080Ti ×3
Development Environments	Python3.8, Pytorch1.8.1 CUDA11.1

2.4. Standard YOLOv8 Model

In order to meet the requirements of precision agriculture for detection and tracking, considering the robustness and accuracy, the standard YOLOv8 has been selected as the baseline model. The architecture of YOLOv8 consists of the backbone network, neck layer, and head layer. The backbone network is responsible for extracting image features across multiple scales. The neck layer fuses these features from the backbone network at each scale. The head layer utilizes three feature maps acquired to predict objects of varying sizes, while YOLOv8 models of different sizes (n, s, m, l, x) share similar structures, their channel depths and the number of convolution modules vary. Figure 4 illustrates the structure of the standard YOLOv8 (YOLOv8n).

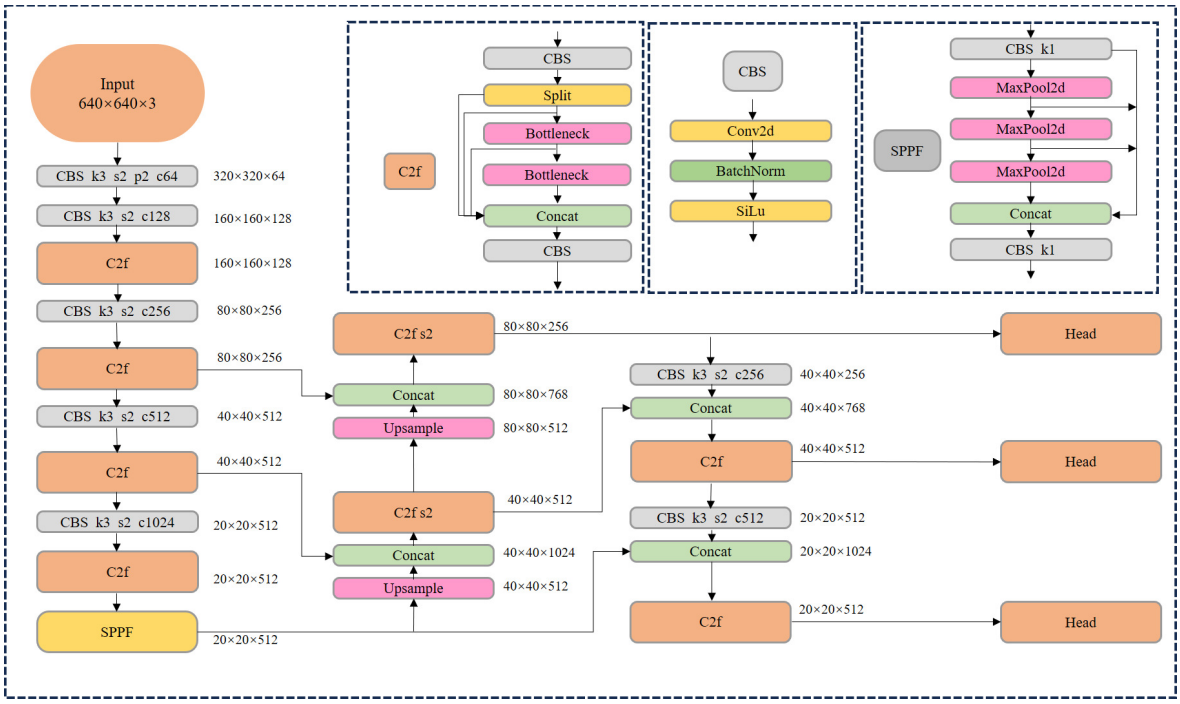

Figure 4. Structure diagram of standard YOLOv8n.

2.5. Improved YOLOv8 Model

This work investigated the corn tracking process and focused on the key issues: significant changes in the size of the corn object and occlusion during tracking. To address these issues, this work proposed YOLOv8MS, which incorporates the Multi-layer Fusion Diffusion Network and separated and enhanced attention module.

2.5.1. Multi-Layer Fusion Diffusion Network

The traditional neck layer in YOLOv8 can only process feature inputs from the same layer. However, during object tracking, the object size often changes significantly, leading to decreased detection accuracy. To address this issue, this work proposed the Multi-layer Fusion Diffusion Network. The MFDN allows features in the neck layer to accept inputs

from three scales and diffuse the output to different layers. The MFDN consists of three parts: the sampling part, fusion part, and diffusion part.

In the sampling part, features from different layers are treated differently to effectively obtain features of varying object scales.

In the fusion part, an inception-like structure was utilized to fuse features, capturing features across multiple scales for improved detection and classification. Additionally, DWConv [21] is used to control parameter count during convolution with a large number of channels.

In the diffusion part, fused features are diffused to different layers of the model. Figure 5 illustrates the network structure of the MFDN.

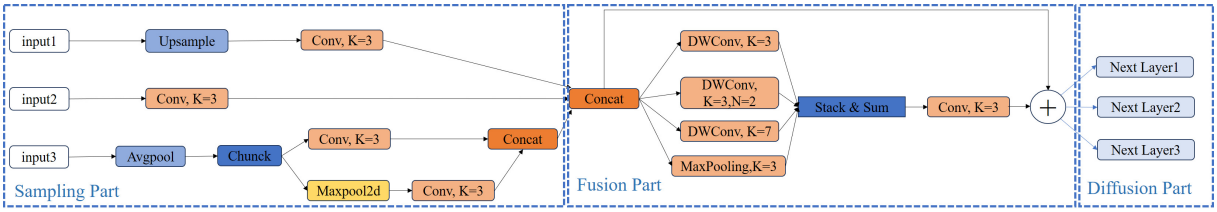

Figure 5. Structure diagram of MFDN.

2.5.2. Separated and Enhancement Attention Module

In the tracking of corn, occlusion by leaves is a common challenge. To address this issue, enable the model to focus more on the object, and reduce background interference, this work introduced the Separated and Enhancement Attention Module.

The SEAM module is designed to handle occlusion efficiently by implementing a multi-head attention network. It includes deep detachable convolution, residual connections, and fully connected networks to strengthen connections between all channels and enhance occlusion handling capabilities. Moreover, the SEAM module utilizes Gaussian Error Linear Units (GELU) [22] to smooth the activation function, addressing the non-differentiability issue at 0 often encountered with Rectified Linear Unit (ReLU).

The SEAM module's design aims to prioritize the object area in the image and reduce emphasis on the background, thereby mitigating occlusion problems. This work integrated the SEAM module into the detection head. Figure 6 illustrates the structure of the SEAM module.

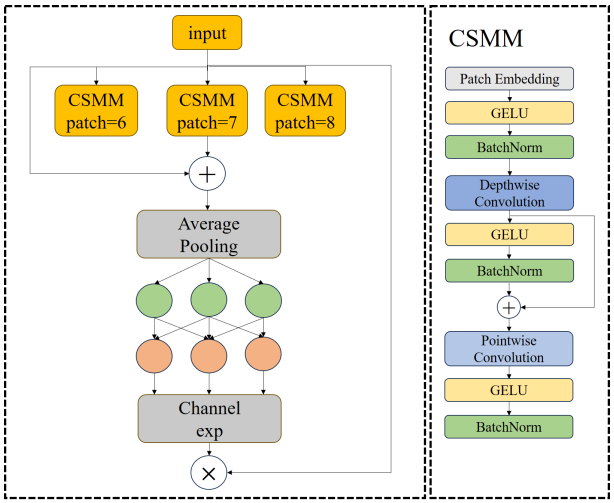

Figure 6. Structure diagram of SEAM module.

2.5.3. YOLOv8MS Model

This work introduces the YOLOv8MS model, which is based on the conventional YOLOv8 model. MFDN replaces the common up-sampling parts of the original network,

facilitating a more effective feature fusion method for improved detection capabilities. Additionally, SEAM is incorporated into the detection head to address the challenge of blade occlusion.

The detailed structure of the YOLOv8MS model is depicted in Figure 7, showing differences from the original YOLOv8 model as illustrated in Figure 4.

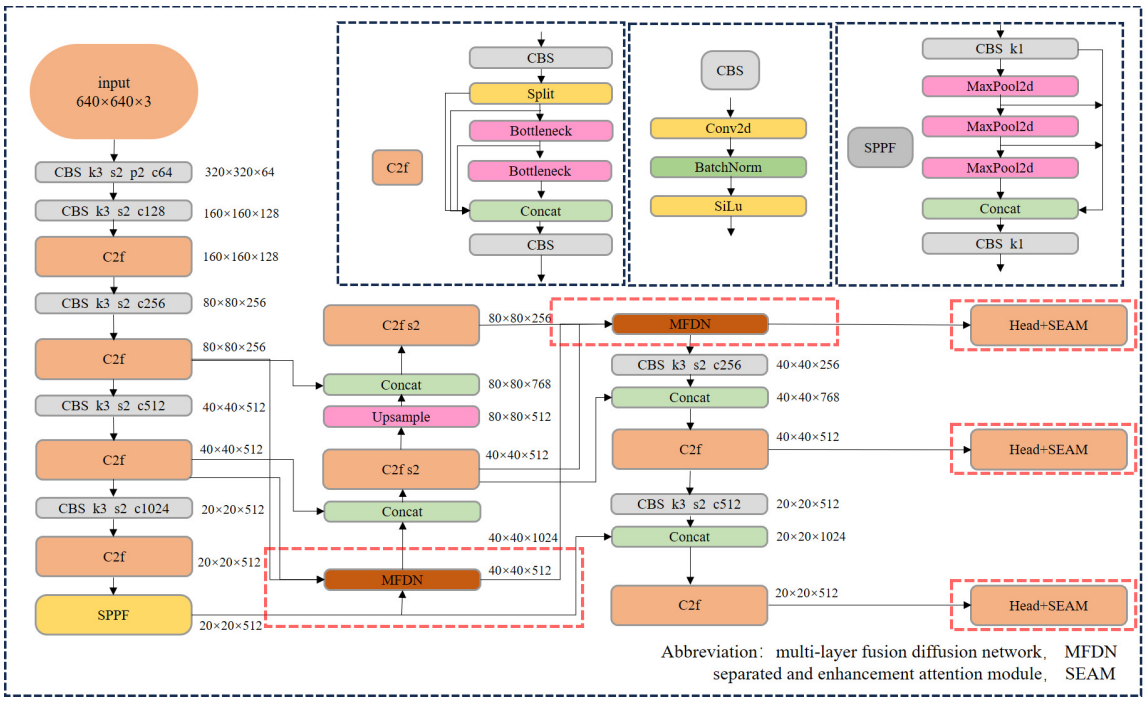

Figure 7. Structure diagram of YOLOv8MS. The red marks mark the changes.

2.6. Model Evaluation Indicators

This work evaluates the algorithm's performance in two aspects: multi-object tracking and object detection.

The multiple object tracking accuracy (MOTA) metric is utilized for multi-object tracking accuracy. MOTA assesses tracking accuracy by considering three types of errors: false positives, false negatives, and identity switches. The MOTA score is computed using Equation (1):

$$\mathrm{MOTA} = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t GT_t} \tag{1}$$

where FN is the number of missed checks in the t frame, FP is the number of false checks, $IDSW$ is the number of identity exchanges, and GT is the actual number of objects.

This work also utilizes identification recall (IDR), identification precision (IDP), and the identification corresponding F1 score (IDF1) [23] to reflect the identity stability of tracking. Equation of IDR, IDP, and IDF1 are summarized as follows:

$$\mathrm{IDF1} = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{2}$$

$$\mathrm{IDR} = \frac{IDTP}{IDTP + IDFN} \tag{3}$$

$$\mathrm{IDP} = \frac{IDTP}{IDTP + IDFP} \tag{4}$$

where $IDTP$, $IDFN$, and $IDFP$ refer to the number of true positive, false negative, and false positive identity assignments, respectively.

This work utilized precision (P), recall (R), and mean average precision mAP to measure the performance of object detection. Specifically, P represents the proportion of positive predictions that are correct as shown in Equation (5), R indicates the proportion of accurate predictions to the total number of all positive samples as shown in Equation (6). AP represents average precision as shown in Equation (7). The mAP represents the mean value of AP under different thresholds as shown in Equation (8). The threshold set ranges from 0.5 to 0.95, and the step size of the threshold is 0.05.

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{P} \tag{6}$$

$$AP = \frac{\frac{TP}{TP+FP}}{N} \tag{7}$$

$$\text{mAP} = \text{mean}\{AP@(0.50 : 0.05 : 0.95)\} \tag{8}$$

3. Result

In Section 3.1, this work compares the results of different YOLOv8 models on the detection dataset. In Section 3.2, ablation experiments of the proposed YOLOv8MS model are conducted on the detection and tracking dataset. In Section 3.3, ablation experiments of the YOLOv8MS on the auxiliary dataset are performed to evaluate the model's performance under challenging conditions.

3.1. Detection Results of YOLOv8 in Different Sizes

Table 4 presents the detection results of YOLOv8 models with different sizes. The results indicate that each YOLOv8 model size achieves good and comparable performance in the detection dataset. Considering the model size, the YOLOv8n model was selected as the benchmark model and further improved upon.

Figure 8 displays the loss change curve of YOLOv8MS and YOLOv8 during training. Both models exhibit similar loss changes, indicating comparable performance in the detection training process.

Table 5 shows the results of the proposed YOLOv8MS model under different optimizers, revealing that the Adam optimizer yields the best detection results.

Table 4. Comparison of object detection results of YOLOv8 models of different sizes.

Model	Precision	Recall	mAP	Parameters
YOLOv8n	99.6%	99.6%	88.6%	3.2M
YOLOv8s	99.5%	99.5%	88.5%	11.2M
YOLOv8m	99.4%	99.6%	88.8%	25.9M
YOLOv8l	99.6%	99.6%	89.0%	43.7M

Table 5. Comparison of the object detection results of YOLOv8MS models using different optimizers.

Model	Optimizer	Precision	Recall	mAP	Parameters
YOLOv8MS	SGD	99.3%	99.4%	88.8%	3.0M
YOLOv8MS	ADAM	99.6%	99.6%	89.6%	3.0M
YOLOv8MS	AdamW	99.3%	99.5%	89.5%	3.0M

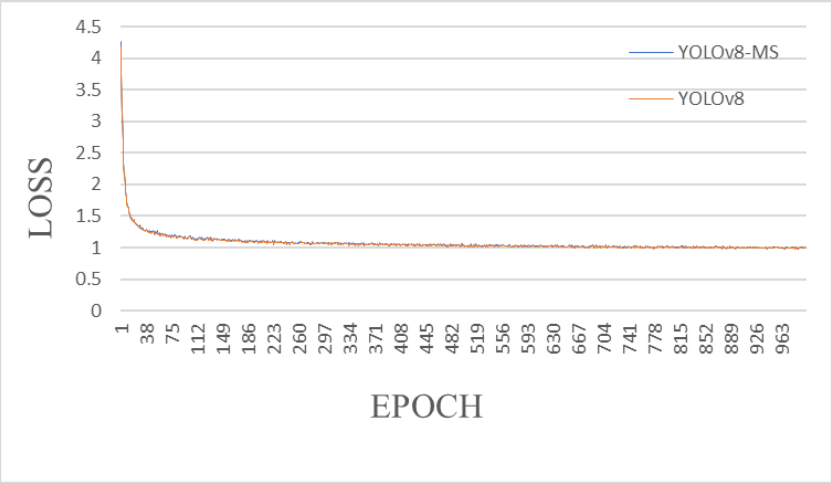

Figure 8. Loss change during the training of YOLOv8MS and YOLOv8.

3.2. Detection and Tracking Results of Ablation Experiments

In this section, ablation experiments were conducted for both detection and tracking tasks. Table 6 shows the detection results of these ablation experiments, demonstrating that all models achieved high detection accuracy.

Table 7 shows the tracking results, indicating that YOLOv8MS outperformed YOLOv8 in tracking accuracy and stability. The results of the ablation experiments further confirm that the SEAM and MFDN modules contribute to improved tracking accuracy and stability.

Table 6. The detection results of ablation experiments.

Model	Precision	Recall	AP50	mAP
YOLOv8	99.5%	99.5%	99.4%	88.6%
YOLOv8 + SEAM	99.5%	99.5%	99.4%	88.5%
YOLOv8 + MFDN	99.5%	99.5%	99.4%	89.1%
YOLOv8MS	99.5%	99.6%	99.5%	89.6%

Table 7. The tracking results of ablation experiments.

Model	MOTA	IDF1	IDR	IDP
YOLOv8	75.5%	84.5%	76.2%	94.7%
YOLOv8 + SEAM	79.1%	86.9%	77.8%	96.5%
YOLOv8 + MFDN	77.8%	85.7%	75.7%	98.7%
YOLOv8MS	81.6%	89.0%	82.3%	97.0%

Figure 9 displays heat maps generated by different models. It is evident from this figure that the proposed YOLOv8MS model effectively reduces interference from the background, assigning more weight to the corn itself during the detection.

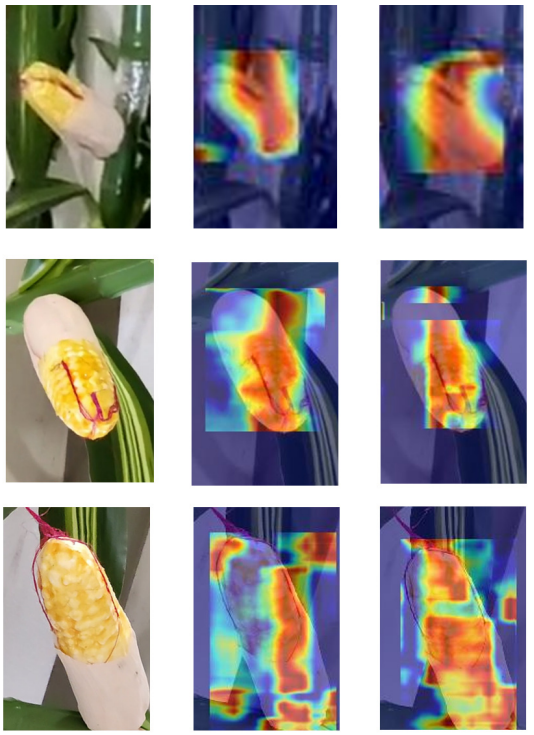

Figure 9. Comparison of attention heatmaps of different models. More red means the model gives more weight and has a greater impact on the results.

3.3. Tracking Results in Challenging Conditions

3.3.1. Results of Auxiliary Dataset of Severe Occlusion

Figure 10 shows images extracted from a video sequence that shows the process from corn being completely shielded by leaves to being exposed and then shielded again. Comparing the results between YOLOv8 and YOLOv8MS under severe occlusion, it is evident that YOLOv8MS can detect objects in more occluded conditions and perform accurate tracking.

Table 8 presents the tracking results of the ablation experiment with YOLOv8MS on the auxiliary dataset of severe occlusion. In this experiment, no identity errors occurred. YOLOv8MS demonstrates superior tracking of corn under heavily occluded conditions. The SEAM module can enhance the model's ability to detect more occluded objects compared to MFDN. Although MFDN has a minor impact on occluded object detection, YOLOv8MS with the SEAM module and MFDN achieved the best results.

Table 8. The results of ablation experiments in the auxiliary dataset under severe occlusion.

Model	MOTA	IDF1	IDR	IDP
YOLOv8	62.5%	76.9%	62.5%	100%
YOLOv8 + SEAM	67.5%	80.6%	67.5%	100%
YOLOv8 + MFDN	58.1%	73.5%	58.1%	100%
YOLOv8MS	92.5%	96.1%	92.5%	100%

Figure 10. The image shows frames grabbed from the video, demonstrating that YOLOv8MS can detect corn more accurately under occlusion. The blue line represents the corn tracking trajectory.

3.3.2. Results of Auxiliary Dataset of Dense Object Presence

The images extracted in the video sequence displayed in Figure 11 show the process of approaching a stack of corn. Comparing YOLOv8 and YOLOv8MS under dense object presence, YOLOv8MS demonstrates more accurate corn detection with fewer errors and missed detections.

Figure 11. The image shows the frames grabbed from the video, demonstrating that YOLOv8MS can detect corn more accurately under dense object presence (the original photos were taken and inputted horizontally). The blue line represents the corn tracking trajectory.

Table 9 presents the tracking results of the ablation experiment with YOLOv8MS on the auxiliary dataset of dense object presence. The results indicate that both the SEAM module and MFDN contribute to improving the model's accuracy and stability of tracking in

dense object presence. The SEAM module has a greater impact, possibly due to significant occlusion within the corn stack. YOLOv8MS with the SEAM module and MFDN achieved the best results.

Table 9. The results of ablation experiments in the auxiliary dataset for dense object presence.

Model	MOTA	IDF1	IDR	IDP
YOLOv8	68.5%	82.1%	73.7%	92.8%
YOLOv8 + SEAM	74.8%	85.5%	75.0%	99.3%
YOLOv8 + MFDN	72.6%	84.2%	73.7%	98.1%
YOLOv8MS	74.9%	86.2%	78.2%	96.0%

3.3.3. Results of Auxiliary Dataset of Varying Viewing Angles

Figure 12 shows images extracted from a video sequence show the process of approaching corn from different angles. The comparison between YOLOv8 and YOLOv8MS under varying viewing angles indicates that both models perform well, although YOLOv8 occasionally assigns incorrect labels.

Table 10 presents the tracking results of the ablation experiment with YOLOv8MS on the auxiliary dataset of varying viewing angles. The results demonstrate that both the SEAM module and MFDN contribute to improving tracking accuracy. YOLOv8MS with the SEAM module and MFDN achieved the best results.

Figure 12. The image shows the frames grabbed from the video, demonstrating that YOLOv8MS can detect corn more accurately under varying viewing angles (the original photos were taken and inputted horizontally). The blue line represents the corn tracking trajectory.

Table 10. The results of ablation experiments in the auxiliary dataset for varying viewing angles.

Model	MOTA	IDF1	IDR	IDP
YOLOv8	94.9%	95.3%	91.3%	99.7%
YOLOv8 + SEAM	96.0%	96.4%	93.4%	99.8%
YOLOv8 + MFDN	95.9%	96.3%	98.9%	100%
YOLOv8MS	97.2%	97.5%	95.4%	99.7%

4. Discussion

This work aims to enhance the accuracy and stability of multiple object tracking of corn for future precision agriculture in planting. To replicate the process of approaching corn and simulate challenging conditions, simulated corn is used to create datasets. The dataset comprises 50 tracking videos, with three auxiliary datasets designed to test the model under challenging conditions such as severe occlusion, dense object presence, and varying viewing angles.

To improve the accuracy and stability of tracking, this work delved into tracking issues and proposed YOLOv8MS based on YOLOv8. The SEAM module is introduced to address occlusion problems and proposed MFDN to fuse features of different sizes, tackling the challenge of varying object sizes during tracking.

Experiments encompassed multiple ablation studies, employing various indicators to demonstrate the enhancements in object detection, tracking accuracy, and tracking stability. This work measured detection accuracy using mAP, tracking accuracy using MOTA, and tracking stability using IDF1, IDR, and IDP.

In the comprehensive tracking test set, YOLOv8MS achieved 89.6% mAP in detection, 81.6% MOTA, 89.0% IDF1, 82.3% IDR, and 97.0% IDP in tracking. This represents a significant improvement over YOLOv8, with respective improvement of 1% mAP, 6.1% MOTA, 4.5% IDF1, 6.1% IDR, and 2.3% IDP.

Under severe occlusion, YOLOv8MS achieved 92.5% MOTA, 96.1% IDF1, 92.5% IDR, and 100% IDP in tracking, outperforming YOLOv8 by 30% MOTA, 19.2% IDF1, 30% IDR, and 0% IDP.

Under dense object presence, YOLOv8MS achieved 74.9% MOTA, 86.2% IDF1, 78.2% IDR, and 96.0% IDP in tracking, outperforming YOLOv8 with respective gains of 6.4% mAP, 4.1% IDF1, 4.5% IDR, and 3.2% IDP.

Under varying viewing angles, YOLOv8MS achieved 97.2% MOTA, 97.5% IDF1, 95.4% IDR, and 99.7% IDP in tracking, outperforming YOLOv8 by 2.3% mAP, 2.2% IDF1, and 4.1% IDR.

Ablation experiments highlighted the efficacy of MFDN and the SEAM module in improving detection accuracy, tracking accuracy, and tracking stability. The SEAM module was particularly effective in scenarios with dense object presence and severe occlusion, while the SEAM module and MFDN showed promising results in varying viewing angles. Combining the SEAM module with MFDN, as proposed in YOLOv8MS, led to even higher accuracy and stability in tracking across all scenarios.

The success of the SEAM module can be attributed to the much occlusion existing in the datasets. In real-world scenarios where variations in corn size are larger, MFDN would play a more substantial role.

While the original YOLOv8 performed as well as YOLOv8MS in detection, YOLOv8MS significantly outperformed it in tracking. YOLOv8 showed tendencies for identity mislabeling due to individual missed detections in the tracking process, reinforcing the suitability of YOLOv8MS for tracking tasks.

Table 11 compares the YOLOv8MS algorithm and previous studies. YOLOv8MS achieved the best MOTA, and MOTAs from previous works were generally lower than the baseline YOLOv8 in this work.

Table 11. Comparison between YOLOv8MS algorithm and previous studies.

Author	Crop	Algorithm	MOTA
Hu et al. [5]	lettuce	YOLO-V5	80.0%
Villacres et al. [7]	Apple	YOLO-V5 and Faster RCNN	59.24% (Average)
Ariza et al. [8]	Grape	CenterNet	66.58%
YOLOv8MS (Ours)	Corn	Keypoint detection	81.6%

Future research plans will explore integrating the proposed YOLOv8MS with unmanned vehicles or drones to track corn and monitor corn growth in farmland. Future

work aims to leverage collected 3D data to conduct research in 3D tracking, which is vital for future advancements.

5. Conclusions

This work focused on improving the accuracy and stability of multiple object tracking for corn, especially under challenging conditions such as severe occlusion, dense object presence, and varying viewing angles. In addition to using the corn tracking dataset, this work also created three additional auxiliary datasets to evaluate the model's performance under these conditions.

YOLOv8n was selected as the benchmark detection model based on experimental results to achieve superior tracking results. Upon analyzing errors, this work considered occlusion and significant changes in object size as major issues during tracking. This proposed YOLOv8MS network will address these issues, incorporating the MFDN and SEAM modules. The MFDN can improve accuracy by accommodating varying corn bounding box sizes. It can fuse context features across multiple scales, leveraging the inception mechanism, and diffuse its output to different layers. The SEAM module includes deep detachable convolution, residual connections, and a fully connected network, and it can effectively manage occlusion by strengthening channel connections.

In the tracking dataset, the experimental results demonstrate notable improvements with YOLOv8MS compared to YOLOv8. In terms of detection accuracy, YOLOv8MS achieved an mAP of 89.6%, outperforming YOLOv8 by 1%. In terms of tracking accuracy, YOLOv8MS achieved an MOTA of 81.6%, outperforming YOLOv8 by 6.1%. In terms of tracking stability, YOLOv8MS achieved IDF1 of 89.0%, IDR of 82.3%, and IDP of 97.0%, outperforming YOLOv8 by 4.5% IDF1, 6.1% IDR, and 2.3% IDP, respectively. In the three auxiliary datasets, YOLOv8MS can achieve better tracking results with improvements in MOTA, IDF1, IDR, and IDP by 12.9%, 8.5%, 3.8%, and 1.1% on average, respectively.

In this work, the proposed YOLOv8MS achieved better object detection accuracy, tracking accuracy, and tracking stability under challenging conditions such as severe occlusion, dense object presence, and varying viewing angles. In the comprehensive corn tracking evaluation, YOLOv8MS can also track corn more accurately and stably.

Author Contributions: Conceptualization, Y.G.; methodology, Y.G.; software, Y.G. Validation, Y.G.; formal analysis, Y.G.; investigation, Y.G. and Z.L.; writing—original draft preparation Y.G.; writing, review, and editing, Y.G.; visualization, Y.G.; supervision, B.L. and L.Z.; funding Acquisition: L.Z. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2133.

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

MFDN	multi-layer fusion diffusion network
SEAM	separated and enhancement attention module
mAP	mean average precision
MOTA	multiple object tracking accuracy
GELU	gaussian error linear units
ReLU	rectified linear unit
IDR	identification recall
IDP	identification precision
IDF1	identification corresponding F1 score

References

1. García-Lara, S.; Serna-Saldivar, S.O. Corn history and culture. In *Corn*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 1–18.
2. Chaivivatrakul, S.; Tang, L.; Dailey, M.N.; Nakarmi, A.D. Automatic morphological trait characterization for corn plants via 3D holographic reconstruction. *Comput. Electron. Agric.* **2014**, *109*, 109–123. [CrossRef]
3. Shi, T. Development and test of automatic corn seedling transplanter. *Trans. Chin. Soc. Agric. Eng.* **2015**, *31*, 23–30.
4. Farjon, G.; Huijun, L.; Edan, Y. Deep-learning-based counting methods, datasets, and applications in agriculture: A review. *Precis. Agric.* **2023**, *24*, 1683–1711. [CrossRef]
5. Hu, N.; Su, D.; Wang, S.; Nyamsuren, P.; Qiao, Y.; Jiang, Y.; Cai, Y. LettuceTrack: Detection and tracking of lettuce for robotic precision spray in agriculture. *Front. Plant Sci.* **2022**, *13*, 1003243. [CrossRef]
6. Yang, H.; Chang, F.; Huang, Y.; Xu, M.; Zhao, Y.; Ma, L.; Su, H. Multi-object tracking using Deep SORT and modified CenterNet in cotton seedling counting. *Comput. Electron. Agric.* **2022**, *202*, 107339. [CrossRef]
7. Villacrés, J.; Viscaino, M.; Delpiano, J.; Vougioukas, S.; Auat Cheein, F. Apple orchard production estimation using deep learning strategies: A comparison of tracking-by-detection algorithms. *Comput. Electron. Agric.* **2023**, *204*, 107513. [CrossRef]
8. Ariza-Sentís, M.; Baja, H.; Vélez, S.; Valente, J. Object detection and tracking on UAV RGB videos for early extraction of grape phenotypic traits. *Comput. Electron. Agric.* **2023**, *211*, 108051. [CrossRef]
9. Wang, L.; Xiang, L.; Tang, L.; Jiang, H. A convolutional neural network-based method for corn stand counting in the field. *Sensors* **2021**, *21*, 507. [CrossRef]
10. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. [CrossRef]
11. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 3464–3468.
12. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. MOT16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
13. ZHANG, K.; REN, H.; WEI, Y.; GONG, J. Multi-target vehicle detection and tracking based on video. In Proceedings of the 2020 Chinese Control Furthermore, Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 3317–3322. [CrossRef]
14. Krupa, K.; Kiran, Y.; Kavana, S.; Gaganakumari, M.; Meghana, R.; Varshana, R. Deep learning-based image extraction. In *Artificial Intelligence and Applications*; Bon View Publishing Pte Ltd.: Singapore, 2022.
15. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3645–3649.
16. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 1–21.
17. Terven, J.; Córdova-Esparza, D.M.; Romero-González, J.A. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.* **2023**, *5*, 1680–1716. [CrossRef]
18. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *arXiv* **2022**, arXiv:2208.02019.
19. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
20. Wang, W. Advanced Auto Labeling Solution with Added Features. 2023. Available online: https://github.com/CVHub520/X-AnyLabeling (accessed on 1 May 2024).
21. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
22. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
23. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 17–35.