JARQ 59 (2), 129-138 (2025) https://www.jircas.go.jp

Performance Evaluation of YOLOv5 for Object Detection in Agricultural Implement Changeover

Van Nang NGUYEN*, Wonjae CHO and Kei TANAKA

Institute of Agricultural Machinery, National Agriculture and Food Research Organization, Tsukuba, Japan

Abstract

The automation of agricultural implement changeovers is crucial for minimizing human intervention in the operation of autonomous farming systems. To ensure system safety and resilience, it is imperative to recognize implements as non-obstacle objects, thereby facilitating the seamless hitching of implements with autonomous tractors. This study presents the initial step in developing a safety function for autonomous implement changeover by assessing the performance of YOLO-based detectors, primarily in terms of precision and speed in detecting target implements and humans. These detectors are trained using transfer learning, employing four YOLOv5 variants (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) and a custom dataset comprising 26,661 labeled images across nine classes of implements and eight classes of obstacles and equipment. The training results show a high average precision (AP) of the detectors, varying from 0.907 to 0.995, for detecting the implements. The mean average precision (mAP@0.5) for detecting all classes ranged from 0.955 to 0.966. Furthermore, testing involving tractor-implement alignments demonstrates the rapid detection of implements and humans by all detectors, with average inference times varying from 7.0 to 20.5 ms. These detectors consistently provide accurate predictions for target objects, with confidence scores (CS) varying from 87.6% to 90.4%. Notably, the detector trained with the medium-variant YOLOv5m is the optimal model with overall performance in terms of both detection speed and accuracy.

Discipline: Agricultural Engineering
Additional key words: confidence score, dataset, detection accuracy, detection speed, safety

Introduction

Modern farmers are embracing smart farming practices to fulfill the increasing demand for agricultural products amid limited resources and a rising population. This involves leveraging autonomous technologies to enhance production efficiency and productivity in a safe and sustainable manner. Japan is making efforts to research and develop autonomous agricultural machinery for smart farming systems, making advancements in three levels (Cho et al. 2021). Level 3, which is the most challenging, realizes fully autonomous operation in driverless conditions, entailing remote monitoring and control of multiple autonomous machines by a single operator.

Consequently, autonomous field-to-field movement on farm roads has been envisioned for remote operations within nearby fields. However, future possibilities, including remote operations on public roads to access more distant fields or the automation of equipment preparation and material supply, are also being explored to enhance the operational efficiency of smart farming systems (Nguyen et al. 2023, Cho et al. 2023).

In the realm of autonomous agricultural operations, the deployment of artificial intelligence (AI) plays a pivotal role in ensuring the safety and efficiency of driverless tasks, particularly in scenarios that involve the operations of autonomous agricultural vehicles. A critical aspect of this autonomy involves the detection and recognition of various obstacles encountered in the agricultural environment, including humans. Driverless vehicles must be equipped with the capability to detect obstacles in proximity and respond promptly to ensure safety.

In the development of an autonomous tractor with a road driving function, an AI model was trained using the

*Corresponding author: nuenb326@affrc.go.jp
Received 26 January 2024; accepted 22 August 2024; J-STAGE Advanced Epub 30 January 2025.
https://doi.org/10.6090/jarq.23J15

YOLO algorithm and public datasets to detect obstacles around the tractor, including five target objects: people, cars, bicycles, motorcycles, and trucks (Cho et al. 2021). The AI model was integrated into the control system of the tractor, enabling it to detect humans and cars and perform emergency stops if necessary. However, in the autonomous implement changeover scenario, the model detects the target implement as an obstacle (Fig. 1). This occurred because the implements were not included in the training dataset, and the detection result could trigger an emergency tractor stop, preventing it from aligning with the implement for hitching.

Consequently, a robust and accurate AI detector that can detect and identify target implements in an outdoor environment is crucial for an autonomous and seamless implement changeover. This study focuses on the performance evaluation of object detectors trained by YOLOv5 with a custom dataset of typical agricultural implements for detecting the target implement and humans. The evaluation was conducted during the process of model training with various YOLOv5 variants and under the deployment of trained detectors in the domain of autonomous implement changeover.

Materials and methods

1. Object detection algorithm YOLOv5

YOLOv5, as a single-stage detector, has the advantages of fast convergence, high precision, high detection speed, and suitability for real-time applications compared to its predecessors (Wang et al. 2023). Therefore, YOLOv5 has been widely applied for the real-time monitoring of crop growth and the detection of diseases, weeds, and obstacles in the agricultural environment (Li, J. et al. 2022, Li, S. et al. 2022, Zhang et al. 2023, Rahman et al. 2023, Chen & Noguchi 2023). These aid farmers in decision-making regarding irrigation and fertilization to prevent the excessive use of pesticides and herbicides, secure the safety of driverless operations, and promote sustainable farming practices.

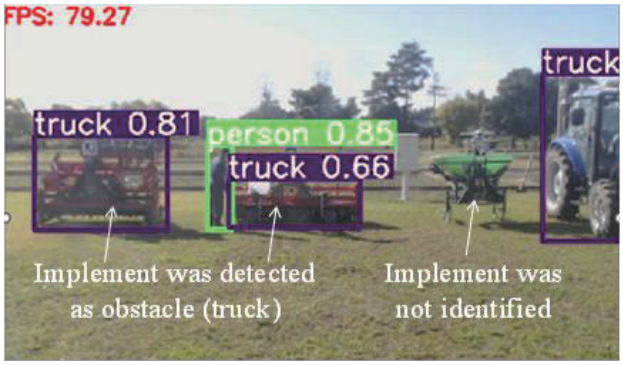

Fig. 1. Implement detection using an AI detector developed for autonomous road driving

For the research in this paper, YOLOv5 was used to train and evaluate the AI detectors utilizing the pre-trained weights of the four YOLOv5 variants (YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) with a custom dataset of agricultural implements. The detector with the best overall performance was determined using several evaluation metrics obtained from model training and testing.

2. Creation of an agricultural implement dataset

A custom dataset was created for nine classes of typical implements (Fig. 2): broad_caster (BC), rotary_tiller (RT), soybean_seeder (SS), wing_harrow (WH), cultivator (CP), plough (TP), power_harow (PH), roll_baler (RB), and fertilizer spreader (DS). Eight other classes were included to represent obstacles, such as persons, agricultural vehicles, and equipment, that may be encountered in the domain of autonomous implement changeover operations. The dataset comprises 26,661 images in 17 categories captured by the authors and collected from the Internet, considering variations within each class, such as machine models, lighting conditions, and backgrounds. All the images were resized to a resolution of 960 × 720 pixels or less.

The image assets were manually annotated with bounding boxes and tags to represent the location and size of the target objects in the images using a Visual Object Tagging Tool (VoTT, Microsoft 2020). The annotation resulted in 48,718 instances of bounding boxes, including 19,998 instances for the implement classes and 28,730 instances for other classes. The labeled images were exported into the Pascal VOC format and converted into YOLOv5 format (text format). The data were divided into a training set with 18,347 images (33,955 instances) and a validation set with 8,314 images (14,763 instances) at a ratio of approximately 7 to 3 (Table 1).

3. Training environment configuration and evaluation metrics

Model training and validation steps were conducted on a desktop computer equipped with an Intel® Core™ i9-12900K CPU, 32 GB RAM, GPU (Graphics Processing Unit): NVIDIA GeForce RTX 3090 Ti with 24 GB VRAM, and Windows 11 Operating System. The training environment was set up using Python (3.10.2, 64-bit) with installed libraries such as PyTorch 2.0.0, YOLOv5, CUDA 11.6, cuDNN, and OpenCV. The key training parameters are listed in Table 2.

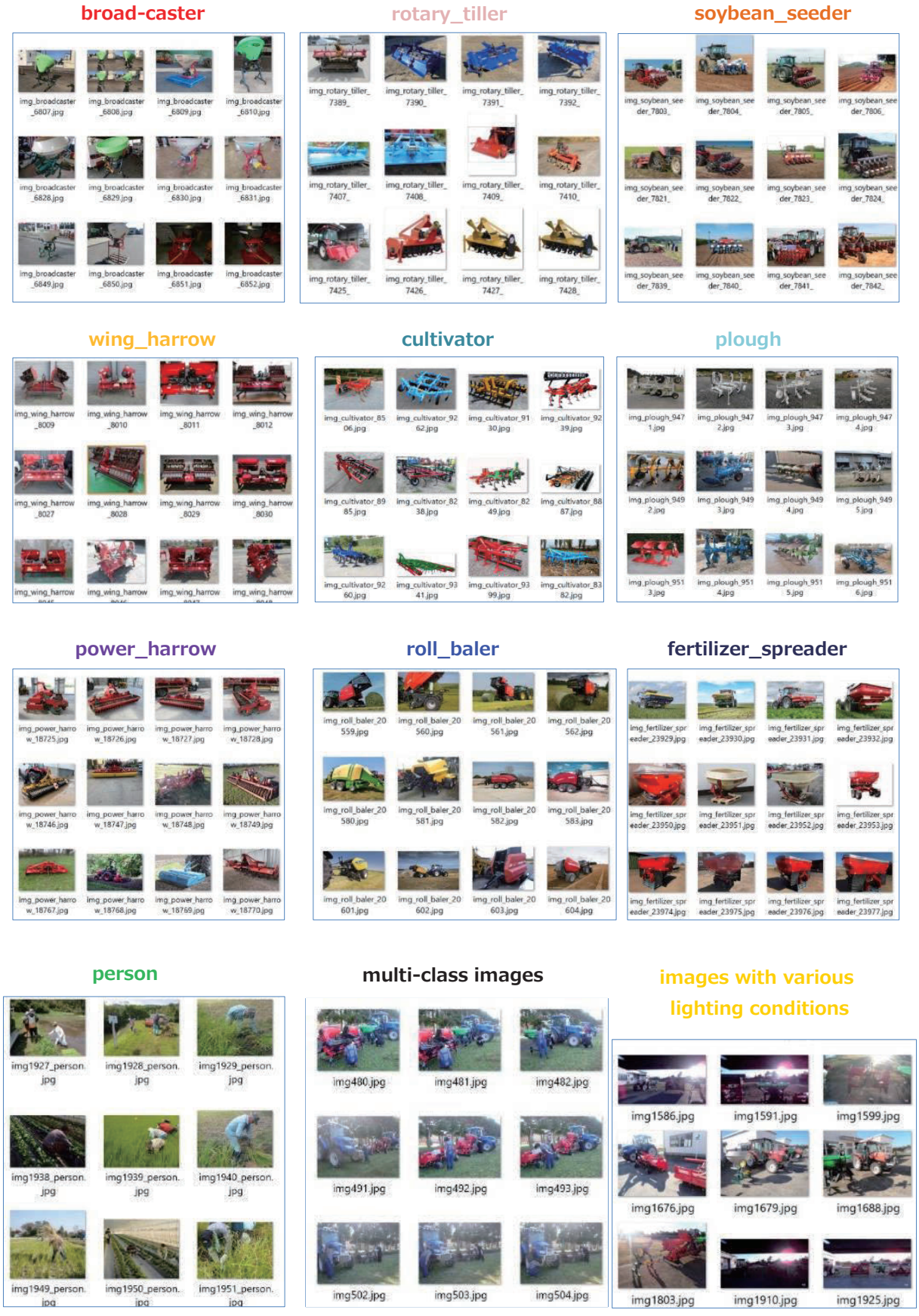

Fig. 2. Sample images of implement and person classes

While precision, recall, and F1-score are useful for minimizing positive detections, false negative detections, and a balance of both, the mean Average Precision (mAP) is an overall metric for comparing different models and assessing their performance across a diverse set of object classes. Accordingly, the performances of the trained detectors were evaluated using the validation set in terms of the maximum values of the metrics (all ranging from 0 to 1) achieved at the point of training convergence.

Table 1. Dataset structure

Class	Number of instances	
	Training set	Validation set
BC	1,937	871
RT	2,032	876
SS	1,423	642
WH	1,668	696
CP	2,163	928
TP	904	383
PH	1,570	661
RB	1,323	558
DS	946	407
person	5,599	3,278
tractor	3,008	853
combine harvester	1,615	710
rice transplanter	1,526	696
truck	1,312	598
hitch	2,361	745
marker	3,278	1,295
bicycle	1,290	566
Total instances	33,955	14,763

4. Testing of trained detectors

All trained detectors were comprehensively evaluated using video sources captured in an outdoor environment under different conditions, including two scenarios of static and moving cameras and two lighting conditions for detecting different target implements and humans (mannequins) in the scene (Table 3 and Fig. 3). The static camera tests were aimed at evaluating the performance of the models in terms of average inference time per image frame, which presents average detection speed (ADS), detection accuracy (DA) calculated by dividing the number of detected frames of target object by the number of captured frames, and average confidence score (ACS). Moving camera tests were used to provide a quantitative understanding of the number of detected objects and the corresponding confidence score (CS) for each class during a tractor-implement alignment following the predefined moving paths of the test camera. The locations of the test implements, mannequins, and static and moving cameras were determined using RTK-GNSS positioning with a smart antenna (A325, Hemisphere), as shown in Figure 3. The test tractor was driven forward and backward at approximately 1.2 km/h during the moving captures.

Four captured videos were fed into the four detectors

Table 2. Major hyper-parameters for model training

Hyper-parameter	YOLOv5 variant			
	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
Pre-trained model size (KB)	14,808	42,811	93,630	174,121
Optimizer	SGD [a]	SGD	SGD	SGD
Image resolution (pixel)	640 × 640	640 × 640	640 × 640	640 × 640
Data augmentation	true	true	true	true
Initial learning rate	0.01	0.01	0.01	0.01
Learning rate factor	0.1	0.1	0.1	0.1
Batch size	16	16	16	16
Number of epochs	300	300	300	300
Patience [b]	20 epochs	20 epochs	20 epochs	20 epochs
IoU [c] threshold	0.2	0.2	0.2	0.2
Weight decay	5×10^{-4}	5×10^{-4}	5×10^{-4}	5×10^{-4}
Training parameters ($\times 10^6$)	7.1	20.9	46.2	86.3

[a] Stochastic Gradient Descent, an iterative optimization algorithm, is used to minimize the loss function during the training of the machine learning model.
[b] Number of epochs with no improvement in the monitored validation loss used to stop training to prevent overfitting.
[c] Intersection over the union, a parameter used to evaluate the performance of object detection by comparing the ground-truth bounding box to the predicted bounding box and calculating their intersection and union areas.

to detect the target implement and human in each image frame with a resolution of 640 × 384 pixels using a laptop computer (Intel® Core™ i7-11800H CPU, 16 GB RAM, GPU: NVIDIA GeForce RTX 3070 with 8 GB VRAM, Windows 10). The thresholds of IoU and confidence score of the detections were set to 0.45 and 60%, respectively. The performances of the detectors under the same object detection conditions were evaluated in terms of ADS, the ratio between the detected objects of each class and the captured frames of each video (DR), and ACS of the detected objects of each class. All testing evaluation metrics were expressed as percentages.

Results and discussion

1. Results of model training and validation

The training and validation results for the four models using the same dataset are summarized in Table 4 and Figure 4, respectively. The number of training epochs decreased with the increasing size of YOLOv5 variants,

Table 3. Test equipment and video capturing conditions

Equipment/lighting condition	Description
Camera	High Dynamic Range (HDR) See3CAM_CU81(e-con Systems) × 1 unit, field of view 120° Horizontal angle, mounting height: approx. 1.5 m Video capture method: static and moving cameras, resolution 1,920 × 1,080 pixel, 30 fps
Tractor	ISEKI TJV85, speed of 1.2 km/h for moving camera
Detection objects	5 units of implements including BC, WH, RT, CP, SS; human: mannequin (1.8 m height)
Lighting condition	Normal light: illumination 106,500 lux, sun elevation 31° - 33°, sun azimuth 138° - 142° Backlight: illumination 80,500 lux, sun elevation 33° - 17°, sun azimuth 218° - 240°

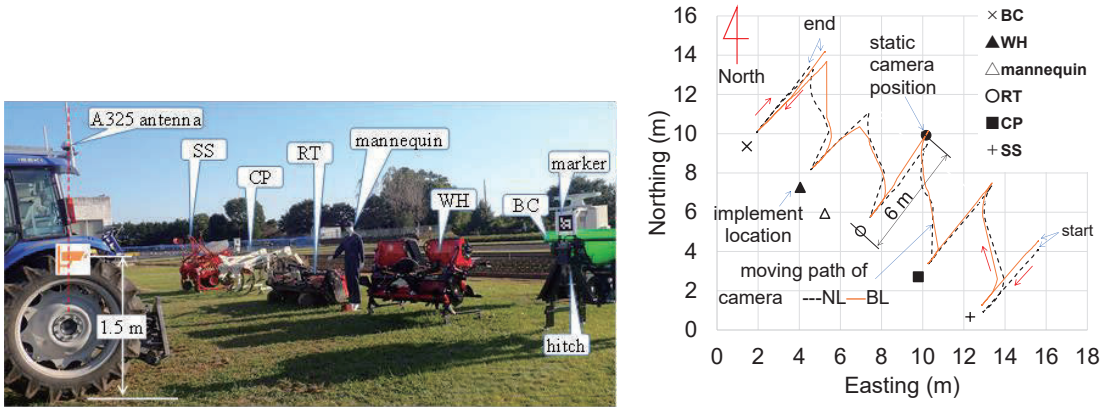

Fig. 3. Outline of test equipment (left) and relative location between implement and camera

Table 4. Training and validation results with different YOLOv5 variants

Training and validation results	YOLOv5 variants			
	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
Training epoch number	244	204	137	127
Model convergence epoch	223rd	183rd	116th	106th
Training time (hour)	9.671	11.829	11.220	16.080
Size of trained detectors (KB)	14,126	41,298	90,781	169,207
Precision	0.948	0.957	0.957	0.959
Recall	0.922	0.937	0.943	0.946
mAP@0.5 of all classes	0.956	0.962	0.963	0.966
F1-score	0.934	0.946	0.947	0.949
mAP@0.5 of implement classes	0.969	0.975	0.975	0.978

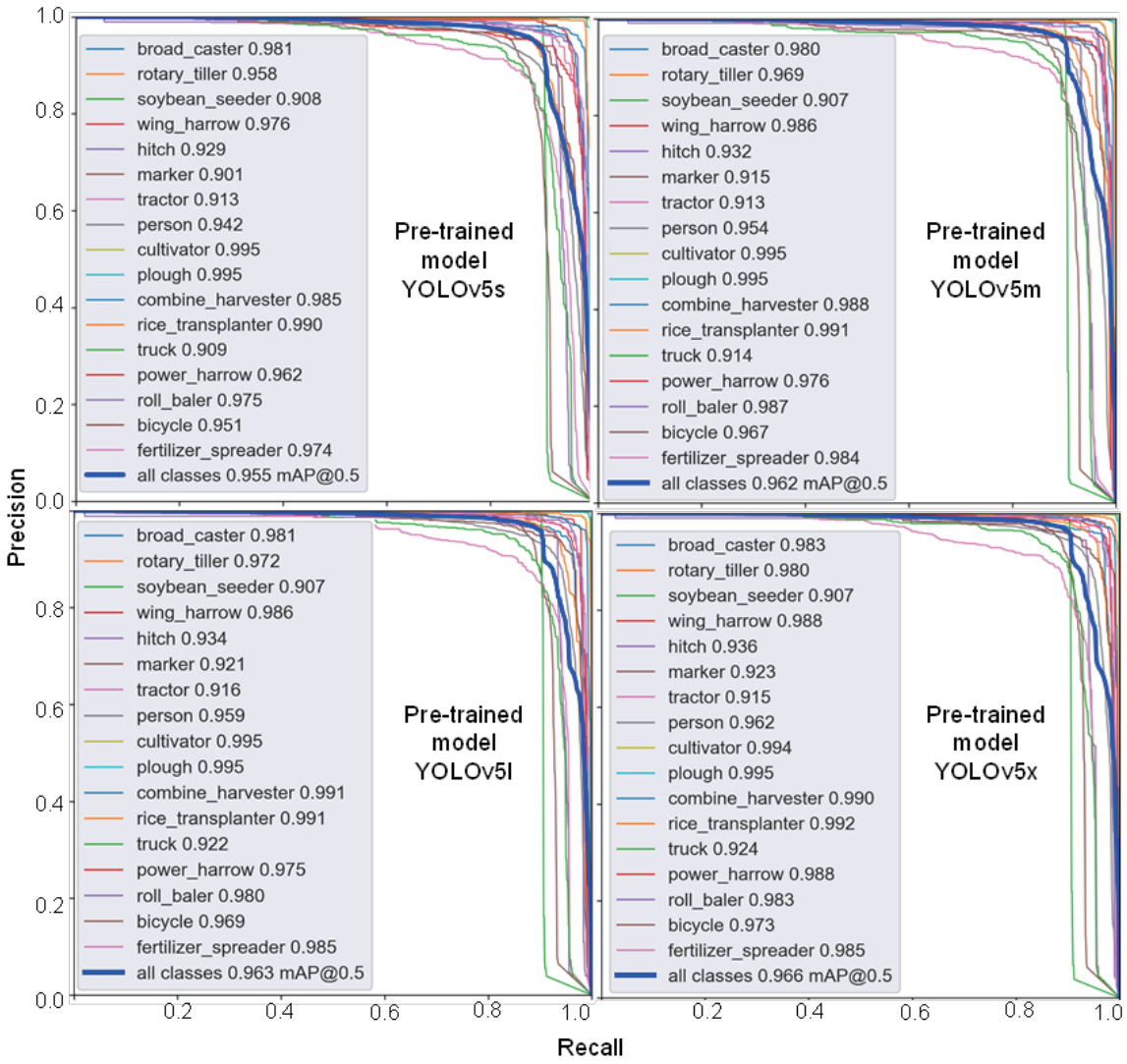

Fig. 4. Precision-recall curves of the trained detectors with different YOLOv5 variants in detecting implements, obstacles, and other equipment

whereas the training time showed the opposite trend. Training with YOLOv5s was stopped after 244 epochs, and the corresponding model converged at the 223rd epoch, taking approximately 9.7 h. Meanwhile, those for other variants were 204 epochs, 183rd epoch, 11.8 h, and 137 epochs, 116th epoch, 11.2 h, and 127 epochs, 106th epoch, 16.1 h, respectively. At the convergence points, the evaluation metrics show a slight improvement as the pre-trained models increase in size. All detectors demonstrated high precision, recall, mAP@0.5 of all classes, and F1-score, varying from 0.948-0.959, 0.922-0.946, 0.956-0.966, and 0.934-0.949, respectively.

The mAP@0.5 of the implement classes varied from 0.969-0.978, indicating that the detectors were capable of detecting the target implements with a high detection accuracy. A similar trend occurred for the person class, with an AP varying from 0.942-0.962. Among the nine implements, despite TP and DS having the smallest training data (less than 1,000 images), their average precision (AP) is extremely high, while SS had the lowest AP with more than 1,400 training images. This suggests that the object shape features of the image datasets of TP and DS are more easily extracted compared to other implements because SS has many machine models in its image dataset. As shown in Figure 4, the larger pre-trained models did not substantially improve the AP of each implement class. Therefore, smaller detectors may be more optimal for detecting implements in the real world, considering their real-time processing capabilities and computing hardware requirements.

2. Testing results

Inference for detecting implements and humans was carried out using four trained weights on the contents of

four videos captured during real-world implement changeover operations. The videos comprised 2,465 and 1,658 frames for the static camera and 6,955 and 6,600 frames for the moving camera under the NL and BL conditions, respectively. Table 5 summarizes the average inference times per frame for the detected objects. The Average Detection Speed (ADS) obtained with the test laptop computer varied from 6.9 to 20.6 ms, depending on the size of the detectors. Lighting conditions had no effect on the ADS obtained with SC and MC. The average ADS for the small, medium, large, and extra-large detectors were 7.0 ± 0.1, 10.3 ± 0.3, 17.4 ± 0.1, and 20.5 ± 0.2 ms, equivalent to object detection speeds of 142.9, 97.1, 68.3, and 48.9 FPS, respectively. Accordingly, all detectors could be utilized for real-time applications, which have a threshold of detection speed from 20 to 30 FPS.

Samples of the detection results for the five target implements and humans under different test scenarios are shown in Figure 5. The detected bounding boxes with corresponding confidence scores were used to calculate the evaluation metrics. For inference with the static camera, DA was calculated by dividing the number of detected objects in each class by the captured frames, as shown in Figure 6. The implements were robustly detected with 100% detection accuracy under both light conditions at distances exceeding 6 m from the camera. In the case of BC, although only half of the implement was captured by the camera, the DA of the BC class was 100% for all inferences. However, in the case of the person class, the DA was significantly reduced to 48.4% and 16.2% with inferences from the small and extra-large detectors, respectively, in the BL condition. This was because the mannequin was smaller than the adjacent target implements, making detection more difficult. In contrast, the medium and large detectors provided high

Table 5. Detection speed of detectors trained with different YOLOv5 variants

Light condition	Camera	Captured frames	Average inference time per frame (ms)			
			YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
NL	static	2,465	6.9	10.1	14.7	20.5
	moving	6,955	7.0	10.2	14.6	20.6
BL	static	1,658	7.2	10.7	14.7	20.5
	moving	6,600	6.9	10.2	14.6	20.2
	Average		7.0	10.3	14.7	20.5

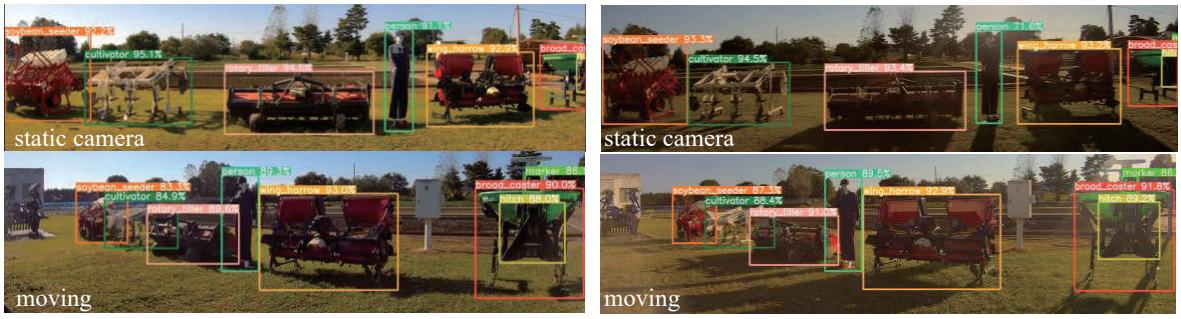

Fig. 5. Samples of detection results in NL condition (left) and BL condition (right)

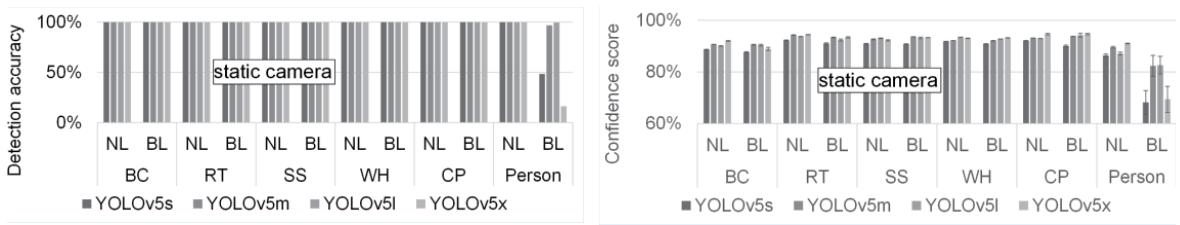

Fig. 6. Detection results of implement and person with a static camera

DA values of 96.9% and 99.6, respectively.

The ACS obtained with four detectors using the static camera varied between 87.6 ± 0.2% to 94.7 ± 0.3% for five target implements and between 68.2 ± 4.6% to 91.0 ± 0.1% for the mannequin. Among the implements, the lowest ACS was observed for BC owing to its location and partial occlusion. The small deviation in the ACS of the implements, together with the high DA, indicated stable implement detection with the detectors deployed in a real environment under different light conditions. A small improvement in the ACS of the implements under the NL condition was observed between the trained weights. However, the medium and large detectors demonstrated higher performance in the detection of humans in the BL condition, with more than 10% improvement in ACS.

Figure 7a illustrates the frequency distribution of the CS obtained with the extra-large detector in the moving camera scenario under NL and BL conditions. Owing to variations in the camera direction and distance, the number of detected bounding boxes and the corresponding CS for each object class varied widely. The values of CS for the implementation and person classes fluctuated between the threshold of 60%, and the maximum values were approximately 95%-97%, respectively. The number of detected objects with a high CS dominated most of the implement and person classes except for BC.

The maximum CS of the target implements, except for RT and BC, was obtained at the closest distance between the camera and the detected objects, as shown in Figure 8. At the final alignment between the tractor and RT, the object features of the implement were reduced, resulting in a lower CS compared to the longer detection distances. In the case of BC, the CS was low during straight backward alignment owing to occlusion by the hitch frame and marker. The consideration of downward-angle cameras for such cases is planned for future evaluations.

Figure 7b demonstrates a small difference in performance among the four detectors under the same light conditions in terms of the number of detected objects and ACS for each class. The number of detected objects for the RT, SS, and CP classes reduced during backlight inference, but no clear differences were observed for WH. In contrast, the number of detected BC increased during BL inference, possibly because of a reduction in the detection of the hitch frame and marker. The total numbers of detected objects under the NL condition were 17,543, 18,320, 18,520, and 18,593 for the small, medium, large, and extra-large detectors, respectively. These values decreased to 16,895, 17,236, 17,412, and 17,486 under BL conditions, respectively. ACS obtained with four detectors was consistent under both NL and BL conditions, varying from 87.6 ± 7.1% to 90.4 ± 4.4%. The high ACS indicates the reliable detection capability of the detectors when implemented in the unseen real operation of the agricultural implement changeover.

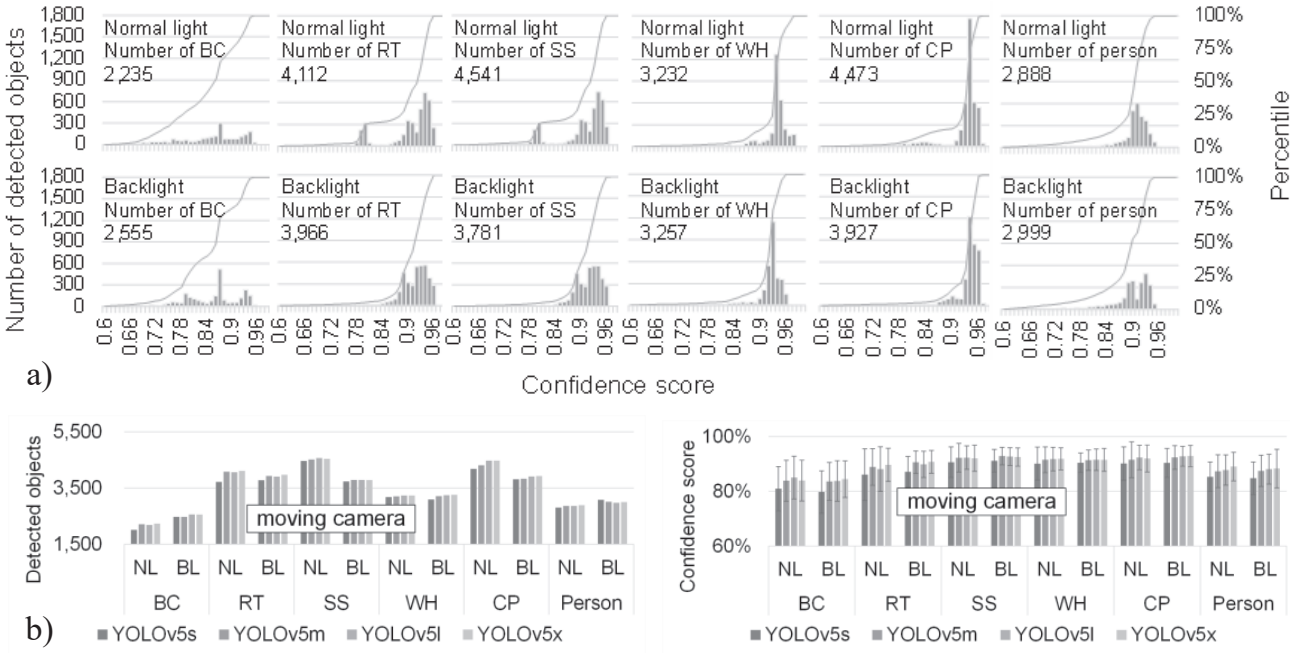

Fig. 7. Detection results of implement and person with moving camera: a) frequency distribution of CS for the extra-large detector; b) Number of detected objects (left) and ACS (right)

Fig. 8. Sample of detection results of implements and person using the extra-large detector during straight backward tractor-implement alignment under BL condition

Conclusion

This study conducted an extensive evaluation of YOLOv5's performance in detecting agricultural implements as non-obstacle objects for seamless tractor-implement alignment during autonomous implement hitching. A custom dataset of nine typical agricultural implements was created. The labeled image data of the agricultural obstacles, including humans, were incorporated into the training and validation sets, which comprised 18,347 and 8,314 images, respectively. The evaluation encompassed both the model training

process, utilizing transfer learning with various YOLOv5 variants, and the deployment of trained detectors for inferences during implement changeover operation under different lighting conditions.

The results of training with the custom dataset demonstrated high accuracy in detecting agricultural implements and obstacles, as the trained AI models exhibited high values of mAP@0.5, varying from 0.956 to 0.966. The test results revealed that all detectors could rapidly predict the detections of five implements under NL and Bl conditions, with average inference time per frame varying from 7.0 to 20.5 ms. At a capturing distance of approximately 6 m, all detectors could detect the implements with a DA of 100% under different lighting conditions, but the performance of the small and extra-large detectors in detecting humans under BL conditions significantly decreased to 48.4 and 16.2%, respectively. However, the medium and large detectors dominated the performance, with a high DA of 96.9 and 99.6%, respectively. Furthermore, the trained detectors accurately predicted the detected objects during the tractor-implement alignments, with confidence scores varying from 87.6% to 90.4%. In conclusion, the medium detector was the optimal model in terms of overall performance, considering the accuracy and speed in detecting agricultural implements and humans under backlight conditions.

Although YOLOv5 brought substantial improvements in accuracy and efficiency compared to its predecessor, its subsequent versions, from YOLOv6 up to YOLOv9, have progressively refined these capabilities, incorporating advanced techniques in feature extraction, model scaling, and post-processing. Particularly, YOLOv9 dynamically adjusts its learning process, improving detection accuracy and robustness in real-world conditions. In the future, the dataset will be expanded by adding more images of equipment and obstacles encountered in the agricultural environment. The detector will be re-trained using the state-of-the-art algorithm to enhance its object detection capabilities.

References

Chen, S. & Noguchi, N. (2023) Remote safety system for a robot tractor using a monocular camera and a YOLO-based method. *Computers and Electronics in Agriculture*, **215**, 108409.

Cho, W. et al. (2021) Development of the integrated control system for an intelligent agricultural vehicle. *Proceeding of the ASABE Annual International Meeting*, Paper No. 2100788.

Cho, W. et al. (2023) Multi-sensor fusion based tractor guidance system for autonomous implement hitching. *J. Jpn. Society of Agric. Machinery*, **85**, 314-323.

Li, J. et al. (2022) An improved YOLOv5-based vegetable disease detection method. *Computers and Electronics in Agriculture*, **202**,107345.

Li, S. et al (2022) A multi-scale cucumber disease detection method in natural scenes based on YOLOv5. *Computers and Electronics in Agriculture*, **202**, 107363.

Microsoft (2020) Visual Object Tagging Tool. https://github.com/microsoft/VoTTcom/yolov5-is-here/. Accessed on 22 September 2023.

Nguyen, V. N. et al. (2023) Development of hitch coupler for autonomous hitching of agricultural implements. *J. Jpn. Society of Agric. Machinery*, **85**, 97-106.

Rahman, A. et al. (2023) Performance evaluation of deep learning object detectors for weed detection for cotton. *Smart Agric. Technol*., **3**, 100126.

Wang, C. et al. (2023) Automatic detection of indoor occupancy based on improved YOLOv5 model. *Neural Comput. & Applic.*, **35**, 2575-2599.

Zhang, Y. et al. (2023) Early weed identification based on deep learning: a review. *Smart Agric. Technol.*, **3**, 100123.

List of Abbreviations

Abbreviation	Definition
ACS	average confidence score
ADS	average detection speed
AP	average precision
BC	broad caster
BL	backlight
CP	cultivator
CS	confidence score
DA	detection accuracy
DR	detection rate
DS	fertiliser spreader
mPA	mean average precision
NL	normal lighting
PH	power harrow
RB	roll baler
RT	rotary tiller
SS	soybean seeder
TP	plough
WH	wing harrow