frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fpls.2022.1091655

TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field

Aichen Wang[1*†], Tao Peng[1†], Huadong Cao[1], Yifei Xu[2,3], Xinhua Wei[1] and Bingbo Cui[1]

[1]Key Laboratory of Modern Agricultural Equipment and Technology (Jiangsu University), Ministry of Education, Zhenjiang, China, [2]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China, [3]Research and Development Department, Nanchang Huiyichen Technology Ltd., Nanchang, China

Check for updates

OPEN ACCESS

EDITED BY
Jun Ni,
Nanjing Agricultural University, China

REVIEWED BY
Abbas Atefi,
California Polytechnic State University, United States
Leiqing Pan,
Nanjing Agricultural University, China

*CORRESPONDENCE
Aichen Wang
acwang@ujs.edu.cn

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Sustainable and Intelligent Phytoprotection, a section of the journal Frontiers in Plant Science

RECEIVED 07 November 2022
ACCEPTED 06 December 2022
PUBLISHED 22 December 2022

CITATION
Wang A, Peng T, Cao H, Xu Y, Wei X and Cui B (2022) TIA-YOLOv5: An improved YOLOv5 network for real-time detection of crop and weed in the field. *Front. Plant Sci.* 13:1091655. doi: 10.3389/fpls.2022.1091655

Introduction: Development of weed and crop detection algorithms provides theoretical support for weed control and becomes an effective tool for the site-specific weed management. For weed and crop object detection tasks in the field, there is often a large difference between the number of weed and crop, resulting in an unbalanced distribution of samples and further posing difficulties for the detection task. In addition, most developed models tend to miss the small weed objects, leading to unsatisfied detection results. To overcome these issues, we proposed a pixel-level synthesization data augmentation method and a TIA-YOLOv5 network for weed and crop detection in the complex field environment.

Methods: The pixel-level synthesization data augmentation method generated synthetic images by pasting weed pixels into original images. In the TIA-YOLOv5, a transformer encoder block was added to the backbone to improve the sensitivity of the model to weeds, a channel feature fusion with involution (CFFI) strategy was proposed for channel feature fusion while reducing information loss, and adaptive spatial feature fusion (ASFF) was introduced for feature fusion of different scales in the prediction head.

Results: Test results with a publicly available sugarbeet dataset showed that the proposed TIA-YOLOv5 network yielded an F1-scoreweed, APweed and mAP@0.5 of 70.0%, 80.8% and 90.0%, respectively, which was 11.8%, 11.3% and 5.9% higher than the baseline YOLOv5 model. And the detection speed reached 20.8 FPS.

Discussion: In this paper, a fast and accurate workflow including a pixel-level synthesization data augmentation method and a TIA-YOLOv5 network was proposed for real-time weed and crop detection in the field. The proposed method improved the detection accuracy and speed, providing very promising detection results.

KEYWORDS

weed detection, deep learning, object detection, YOLO, precision agriculture, site-specific weed management

1 Introduction

During the process of crop growth, weeds appear randomly in the field and compete with crops for water, nutrients and sunlight, leading to a negative effect on crop yield and quality (Zhang, 2003; Lee et al., 2010). Research shows that weed competition may cause crop yield loss as high as 34% (McCarthy et al., 2010; Gao et al., 2018). Weed control has become one of the most important tasks in modern agricultural. Chemical and mechanical weeding campaigns are the two main means adopted at present. However, the overuse of herbicides by chemical weeding operations has resulted in serious environmental pollution problems. Mechanical weeding operations by tillage or cultivation of soil have been widely used for crops planted in rows, but they could hardly remove intra-row weed without assistance from a target detection module and may cause severe crop damage. In order to solve these problems, the concept of site-specific weed management (SSWM) was introduced, which could provide accurate target information for both the herbicide spraying system and mechanical weeding equipment (Lottes et al., 2018).

The key of SSWM is the rapid and accurate detection of target. For chemical weeding, the detection target is weeds, while for mechanical weeding, the detection target is crops for reducing crop damage rate. Therefore, rapid and accurate detection of both crops and weeds is very important for SSWM operations. At present, visible-near infrared (Vis-NIR) spectroscopy and machine vision are two main techniques used for crop and weed detection in the field. The Vis-NIR spectroscopy technique shows absolute advantage in detection speed, but it could hardly distinguish the weak spectral difference between crops and weeds in complicated outdoor environment. In addition, the detection region of Vis-NIR detector is limited, making this technique not suitable for crop and weed discrimination in the field (Wang et al., 2019). Compared with the Vis-NIR spectroscopy, machine vision could acquire information in a large area and provide accurate location information in the field of view. The image processing methods for machine vision can be grouped into conventional hand-crafted feature-based method and deep learning-based method. The conventional image processing method trains machine learning models based on color, texture, shape and other hand-crafted features. It has a simple model training process, but the model generalization ability and adaptability are low, hindering its practical applications under different circumstances. Moreover, the image preprocessing of the conventional method is cumbersome (Wang et al., 2022c). In recent years, with the rapid development of convolutional neural network (CNN) for image and video processing, deep learning-based image processing method has been widely studied and applied in the field of agricultural engineering (Wang et al., 2020a; Hasan et al., 2021). Deep learning-based object detector shows great advantages in target recognition, positioning and category determination. Jiang et al. (2020) established a grap convolutional network (GCN) map by extracting weed feature map and Euclidean distance through a CNN network, and enriched image features by using the GCN map. The recognition accuracy of GCN-ResNet-101 model on four weed datasets reached 97.80%, 99.37%, 98.93% and 96.51%, respectively, but the model convergence process was slow, and the model training was difficult. Furthermore, the network structure cannot be set too deep due to the limitation of the GCN network, otherwise it would cause the vanishing gradient problem. Kim and Park (2022) trained a multi-task semantic segmentation-convolutional neural network (MTS-CNN) model based on U-Net for the semantic segmentation of weeds and crops. Considering the large difference in the loss function between crops and weeds, they designed the cross-entropy loss and dice loss models between weeds, crops and both (weeds and crops) in the loss function stage to increase the stability of the network. The mean intersection over union (MIoU) trained on three public datasets was 91.61%, 83.72% and 82.60%, respectively. However, the model improvement was not based on the characteristics of specific objects (crops and weeds), resulting in poor generalization of the model. Peng et al. (2022) proposed an improved RetinaNet model to detect weeds among rice crops. The convolution structure was modified to reduce the loss of semantic information, the Efficient Retina Head (Lin et al., 2017) was designed in the head network to reduce memory consumption and inference time, and the regression loss function was designed by combining the Smooth loss (Girshick, 2015) and generalized intersection over union (GIoU) loss (Rezatofighi et al., 2019). Results showed that the average weed recognition accuracy of the model was 94.1%, which was 5.5% and 9.9% higher than the average recognition accuracy of the baseline network RetinaNet and YOLOv3, respectively. Although the model had some improvements in the prediction head, it did not make full use of the rich semantic information extracted from the backbone network, which cut the correlation between the feature maps. As a result, it was unfavorable for the practical application of the model.

Therefore, the overall objective of this work was to develop a fast and accurate model for weed and crop object detection in the field. Specifically, a pixel-level synthesization data augmentation method was proposed to deal with the problem of unbalanced sample distribution of weed and crop. An improved YOLOv5 network named the TIA-YOLOv5 was proposed, in which a transformer encoder block was added to the backbone to improve the sensitivity of the model to weeds, a channel feature fusion with involution (CFFI) strategy was proposed for channel feature fusion while reducing information loss, and adaptive spatial feature fusion (ASFF) was introduced for feature fusion of different scales in the prediction head. Lastly, the effectiveness of the proposed pixel-level synthesization data augmentation method and TIA-YOLOv5 network was tested with a publicly available sugarbeet dataset for sugarbeet and weed detection.

2 Materials and methods

2.1 Proposed method

YOLO series networks have been widely used in the field of agriculture for object detection (Wang et al., 2022b; Wang et al., 2022e). By now, the YOLO series models have developed into the seventh version, in which YOLOv5 is the most widely used for object detection. YOLOv5 is mainly composed of a (Cross-stage-Partial-connections) CSP-Darknet53 (Bochkovskiy et al., 2020) as the backbone, a path aggregation network (PANet) as the neck and a YOLO as the prediction head (Glenn, 2020). To improve the detection accuracy of crop and weed in the field, the YOLOv5s was selected as the baseline network in this work. The YOLOv5s is a light object detector with high performance, making it suitable for real-time object detection and easy to deploy on an edge computing platform in the field.

The architecture of the proposed TIA-YOLOv5 model is depicted in Figure 1. The model includes a backbone, a neck and a prediction head. To deal with the problems caused by occlusion, high density and sharp scale change in the current weed object detection scenario, the TIA-YOLOv5 network has three improvements compared with the baseline YOLOv5 model. Firstly, a transformer encoder block (Zhu et al., 2021) was added into the backbone to increase the feature extraction capability of the backbone network. Secondly, a channel feature fusion with involution (Li et al., 2021a) (CFFI) module was designed to reduce the loss of semantic information caused by convolution operations in the feature fusion stage and make full use of the rich information of the feature map at the end of the backbone. And thirdly, adaptively spatial feature fusion

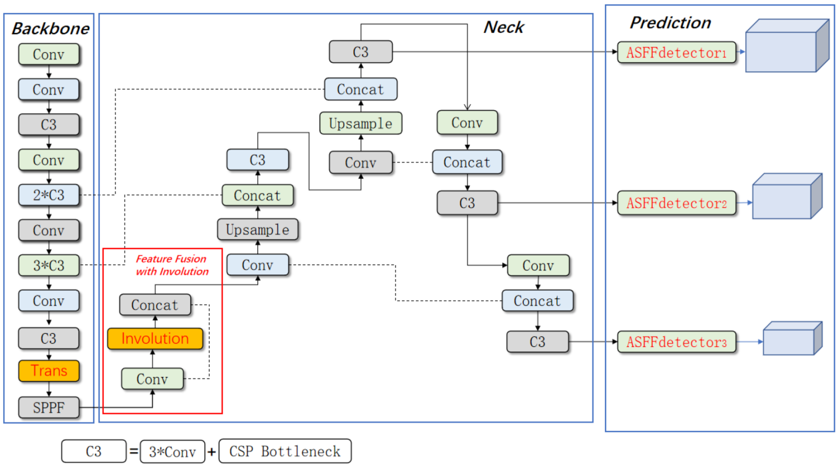

FIGURE 1
Architecture of the proposed TIA-YOLOv5 network. (* means multiply).

(Liu et al., 2019) (ASFF) was introduced to enhance the sensitivity of the prediction head to crops and weeds.

2.1.1 Transformer encoder

As the backbone of YOLOv5s, the CSP-Darknet53 is mainly composed of CSP bottleneck blocks, which is based on residual mechanism. The combination of transformer (Dosovitskiy et al., 2020) and CSP module is getting more attention by researchers recently due to its strong feature extraction capability. In the backbone, the transformer encoder block was deployed to replace the bottleneck of CSP-Darknet53. The transformer encoder block can get small target information better in the global information through self-attention mechanism, and has better detection performance in high-density crop images. The structure of the transformer encoder block is shown in Figure 2. It is mainly composed of a multi-head attention sub-layer and a multilayer perceptron (MLP) feed-forward neural network sub-layer with residual connections between them. In this work, the transformer encoder block was placed at the backend of the backbone, because the output of the backbone network is a low-resolution feature map, which will reduce the computational load of the transformer encoder block (Zhu et al., 2021).

2.1.2 Channel feature fusion with involution (CFFI)

Path aggregation network (PANet) (Liu et al., 2018), which pools features from pyramid levels and fuses features of different scales, is widely used in object detection networks including YOLOv5. As the high-level feature map of the backbone, feature {C5} contains rich semantic information. To integrate with the mapping of the feature {C4} in the backbone, 1 × 1 convolution layers were adopted to reduce the channel number of feature {C5}, through which the calculation efficiency was significantly improved. However, the reduction of channel number would inevitably result in serious information loss (Luo et al., 2022). To reduce this information loss, the CFFI was added between the backbone and feature pyramid, which enabled the semantic information to be fully utilized. Based on this concept, in order to take full advantage of the rich features in the high-level channels of the backbone and improve the performance of the PANet, involution operation (Li et al., 2021a) was adopted. Involution is characterized by light weight, high efficiency and flexibility, and has achieved good results in various visual tasks. Different from convolution operation, involution has channel invariance and spatial specificity, enabling it to overcome the difficulty of modeling long-range interactions (Li et al., 2021a).

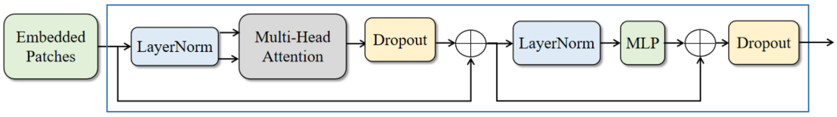

FIGURE 2
Structure of transformer encoder block (Zhu et al., 2021).

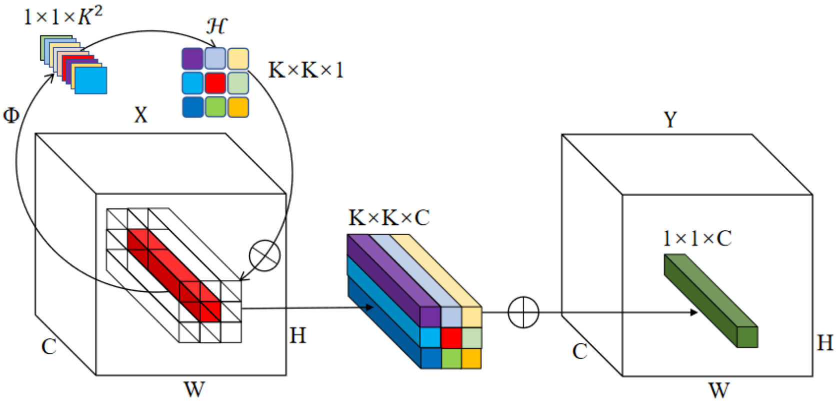

FIGURE 3
Schematic illustration of involution (Li et al., 2021a).

The formula of the involution kernel is $\mathcal{H}_{i,j} \in \mathbb{R}^{K \times K \times 1}$, where $\mathcal{H}_{i,j}$ generates according to the function Φ at a pixel (i, j), K is the kernel size of involution. The structure of involution is shown in Figure 3. In involution, the size of the input feature map directly determines the size of involution kernel, therefore the size of the kernel and the input feature map automatically aligns in the spatial dimension. The advantage of this is that the involution kernel can adaptively allocate the weights over different positions (Li et al., 2021a).

It was observed that the channel number of the feature map {C5} can maximize the performance of involution. Therefore, the CFFI was introduced to directly reuse the rich semantical information of the feature {C5} (Luo et al., 2022). Firstly, a 1 × 1 convolution with channel number of 768 was used to reduce the channel number of the feature {C5}. Then an involution with kernel size of 1 × 1 and channel number of 768 was used to aggregate spatial information. Finally, a concat operation was conducted to connect the resulted feature maps of convolution and involution, as shown in Figure 4. Through these operations of CFFI, the channel number of feature maps was reduced, as well as the loss of semantic information caused by direct convolution. Therefore, the CFFI functioned as a bridge between the backbone and neck, which enhanced the representation ability of the feature pyramid.

2.1.3 Multiscale feature fusion

YOLOv5 is a single-stage detector. The main problem of YOLOv5 is the inconsistency between feature maps of different scales (Liu et al., 2019). For weed detection task, there are large scale differences between crops and weeds, resulting in inaccurate positioning and identification of objects. In order to make full use of the multi-layer features output from the neck, the adaptively spatial feature fusion (ASFF) (Liu et al., 2019) was introduced. Unlike the bi-directional cross-scale connections and weighted feature fusion (BIFPN) and learning scalable feature pyramid architecture (NAS-FPN) that use cascaded multi-level feature fusion, features of different scales were rescaled and adaptively fused in ASFF to filter out the inconsistency during training. The structure of the ASFF is shown in Figure 5.

There are three resolutions of feature maps in YOLOv5, $\boldsymbol{P_2}$, $\boldsymbol{P_3}$, and $\boldsymbol{P_4}$. For feature map of $P_i(i \in \{2, 3, 4\})$, the resolution of other $P_n(n \neq i)$ feature maps was adjusted so that all feature maps have the same size. The ASFF modifies the upsampling and downsampling strategies for the three feature maps. For up-sampling, the 1 × 1 convolution was used to reduce the channel number of feature maps, then interpolation was conducted to improve the resolution of feature maps. Two downsampling strategies were used; For the feature map of 1/2 scale, the 3 × 3 convolution with a stride of 2 was used to modify the channel size and resolution simultaneously, and for the feature map of 1/4 scale, the maximum pooling with a stride of 2 was added before the convolution layers with a stride of 2. The last step of ASFF was to adaptively learn the spatial weights of feature map fusion under multiple scales for adaptive fusion (Liu et al., 2019). The formula of ASFF is as Eq. (1).

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1\to l} + \beta_{ij}^l \cdot x_{ij}^{2\to l} + \gamma_{ij}^l \cdot x_{ij}^{3\to l} \tag{1}$$

Where y_{ij}^l is the (i,j)-th vetor of the output feature map y^l between channels, and $x_{ij}^{n\to l}$ represents the feature vector at the position (i, j) on the feature map adjusted from the n level to the level. α_{ij}^l, β_{ij}^l, γ_{ij}^l, are the spatial weights corresponding to the three levels of feature maps, which were obtained by network adaptive learning. α_{ij}^l, β_{ij}^l and γ_{ij}^l are subjected to the Eq. (2).

$$\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$$

$$\alpha_{ij}^l, \quad \beta_{ij}^l, \quad \gamma_{ij}^l \in \quad [0, 1] \tag{2}$$

$$\alpha_{ij}^l = \frac{e^{\lambda_{a_{ij}}^l}}{e^{\lambda_{a_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}}$$

By adopting this strategy, adaptive fusion was carried out on each feature scale, which greatly reduced the instability of the model caused by the scale changes between crops and weeds.

2.2 Dataset preparation

2.2.1 Image dataset

In this work, a publicly available sugarbeet image dataset (Chebrolu et al., 2017) was used. The images were captured with a readily available agricultural robotic platform, BoniRob, on a

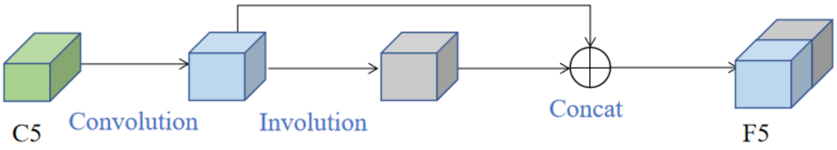

FIGURE 4
Channel feature fusion with involution.

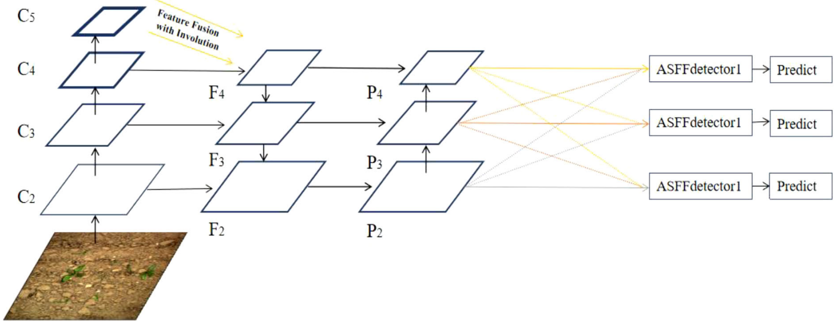

FIGURE 5
Schematic diagram of adaptively spatial feature fusion (ASFF).

sugar beet farm near Bonn in Germany in spring 2016. All images are in RGB format and the resolution is 1296 pixel × 966 pixel. The dataset includes images of sugar beet at seedling and growth stages as well as weeds in the field (Chebrolu et al., 2017). In this study, a total of 4500 images were selected as our raw dataset. LabelImg (https://github.com/tzutalin/labelImg) was used to label crop and weeds in the images and provide corresponding label files for model training.

2.2.2 Image preprocessing

The balance of samples determines the robustness of the trained models. During the process of labeling the sample data, it was found that the sample number of weeds and sugarbeet in the public image dataset was seriously unbalanced, with the ratio of sugarbeet to weeds was about 17:3. To increase the number of weed samples, a pixel-level synthesization data augmentation method was designed. Six kinds of weeds with different sizes and shapes were extracted from original images, then these weed pixels were inserted randomly into images to synthesize new images containing more weed objects. The data augmentation process is shown in Figure 6. Sub-images (Figure 6A) that only contained weeds and background were cropped from the original images. The background and weeds in the sub-image were then segmented using the region growth method (Angelina et al., 2012). The extracted weed pixels are shown in Figure 6B. These extracted weed pixels were merged into original images randomly to obtain new images (Figure 6C) containing more weed objects. After data augmentation, the number of total images reached 5536, and the sample ratio of sugarbeet to weeds was 17:9.

The enhanced dataset containing 4100 original images and 1436 synthetic images was divided into a training set, a validation set and a test set with a ratio of 8:1:1. Since the pixel-level synthesization data enhancement method is a process of data expansion based on the original images, to avoid the repetition of weeds and crops in the training set and test set, the

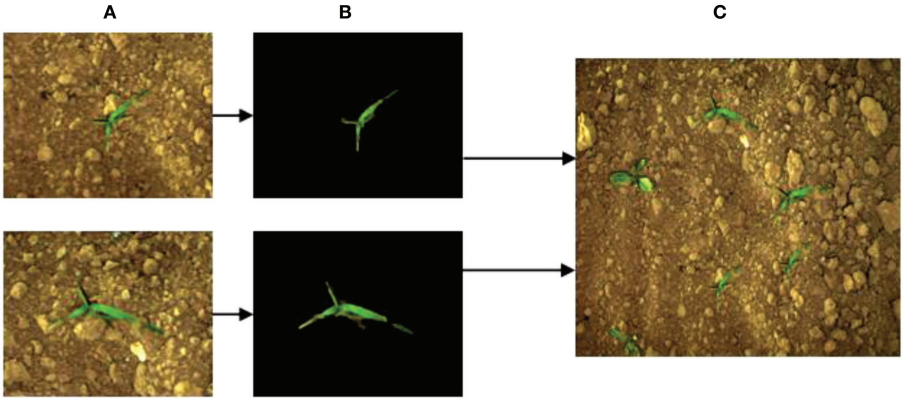

FIGURE 6
Image augmentation process for increasing the number of weed objects. **(A)** Sub-image containing only weeds and background, **(B)** Extracted weed pixels, **(C)** Synthetic new images containing more weed objects.

obtained synthetic images and corresponding original images that were used for date augmentation were only used for model training. As a result, the resulting training, validation and test set contained 4395, 500 and 641 images, respectively.

2.3 Experiment design

The model training and testing tasks in this work were conducted on a PC equipped with a 11th Gen Intel i9-11900K CPU and a NVIDIA GTX3080Ti GPU. Pytorch 1.10.2 framework was used to build the networks. For parameter settings, the pretrained weights of YOLOv5s on the COCO dataset was loaded as initial weights. The input image size was 1296 × 966, the training epochs was set to 100, and the batch size was set to 8. To verify the effect of the data augmentation method on the proposed TIA-YOLOv5, conventional data augmentation method including random translation, rotation and scaling was also used for comparison. Six groups of ablation experiments were conducted to analyze the effects of each module of the TIA-YOLOv5 network. Finally, the TIA-YOLOv5 were also compared with state-of-the-art object detection models including SSD (Liu et al., 2016), Faster RCNN (Girshick, 2015), YOLOv3 (Redmon and Farhadi, 2018), YOLOv4 and YOLOv7 (Wang et al., 2022a) networks.

2.4 Evaluation metrics

Precision, recall, average precision (AP), mean average recognition accuracy (mAP@0.5 and mAP@0.5:0.95), F_1-score and processing speed in terms of frame per second (FPS) were used to evaluate the model performance. The mAP is the average of the AP calculated for all the classes, where N is the number of categories. The mAP@0.5 means that the mAP calculated at intersection over union (IoU) threshold of 0.5. The mAP@ 0.5:0.95 means the average mAP over IoU thresholds from 0.5 to 0.95 with an interval of 0.05. The formulas for precision, recall, F_1-score, AP and mAP are as follows.

$$Precision = \frac{True\ \ Positive}{True\ \ Positive + False\ \ Positive} \times 100\,\% \qquad (3)$$

$$Recall = \frac{True\ \ Positive}{rue\ \ Positive + False\ \ Negative} \times 100\,\% \qquad (4)$$

$$F1 - score = 2 \times \frac{Percision \times Recall}{Percision + Recall} \times 100\,\% \qquad (5)$$

$$AP = \int_0^1 Percision(Recall)d(Recall) \times 100\,\% \qquad (6)$$

$$mAP = \frac{\sum_1^N \int_0^1 Percision(Recall)d(Recall)}{N} \times 100\,\% \qquad (7)$$

3 Results and discussion

3.1 Performance of dataset augmentation

To evaluate the performance of the proposed pixel-level synthesization dataset augmentation method, commonly used conventional dataset augmentation methods including random translation, rotation and scaling were used for comparison. The number of samples in the training, validation and test set for each method was set the same. The results are shown in Figure 7. Because the precision, recall, F_1-score, and AP values for detecting sugarbeet were all above 90% for all models, they were not listed and discussed as performance indicators. The YOLOv5 model without data augmentation performed slightly better than the 'YOLOv5+conventional augmentation method' model. This is because the 'YOLOv5+conventional augmentation method' replaced 1436 original images with corresponding processed images by the conventional data augmentation method to keep the same number of samples in the training set. The proposed pixel-level synthesization method were superior to the conventional method in terms of all the six indicators. The conventional data augmentation method could enhance the dataset to some extent, however, the lack of weed samples that results in the unbalanced data distribution is the key problem in this work, which cannot be solved by the conventional augmentation method. The pixel-level synthesization method was designed to increase the number of

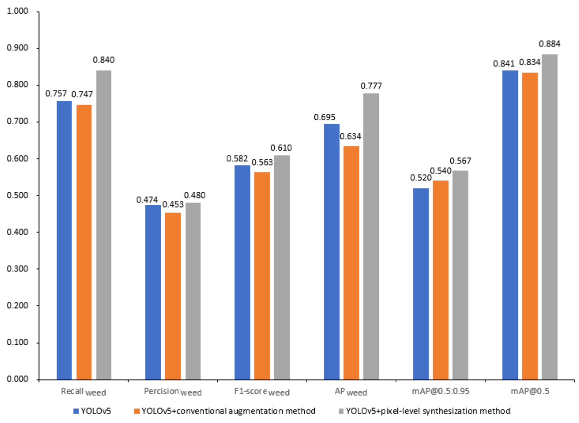

FIGURE 7
Performance of data augmentation methods.

weed samples in the dataset, further to alleviate the unbalanced phenomenon of data distribution. From Figure 7 it can be observed that the F_1-score$_{weed}$, AP$_{weed}$ and mAP@0.5 were improved to 61.0%, 77.7% and 88.4%, with 4.7%, 14.3% and 5.0% absolute increase compared with the conventional data augmentation method, respectively, validating the effect of the pixel-level synthesization method.

3.2 Ablation study

Ablation study was conducted to evaluate the effect of different modules of the proposed TIA-YOLOv5 network. The results of ablation study are shown in Table 1, from which it can be seen that the Recall$_{weed}$, F_1-score$_{weed}$, AP$_{weed}$, mAP@0.5 and mAP@0.5:0.95 of the 'YOLOv5+Transformer' model increased by 11.3%, 3.6%, 8.3%, 4.4% and 5.9%, respectively, compared with the baseline YOLOv5 network. These improvements indicate that the transformer encoder block could improve the sensing ability of the YOLOv5 network for weeds, making it easier to capture the weed objects from the complicated context information. In addition, the calculation cost of the transformer encoder block was less than that of the original bottleneck module of the YOLOv5 network, which improved the FPS of the model. The combination of the pixel-level synthesization augmentation method and transformer encoder block reduced the additional computer resource consumption caused by the increase of samples. Although adding the CFFI module alone cannot improve the model performance on identifying crops and weeds significantly, the Recall$_{weed}$, F_1-score$_{weed}$, AP$_{weed}$ and mAP@0.5 of the 'YOLOv5+Transformer+CFFI' model reached 86.0%, 62.5%, 79.8% and 89.4%, with an absolute increase of 10.3%, 4.3%, 10.3% and 5.3% compared with the baseline network, when combining the transformer encoder block with the CFFI module. This is because the CFFI module could make full use of the rich context information captured by the transformer encoder block, realizing the reuse of the information extracted from the backbone network and further providing rich front-end information for feature fusion layers in the neck (Zhu et al., 2021). The ASFF module could further optimize the network by applying multi-scale feature fusion in the prediction head. In this work, the feature maps of were fused to reduce the instability of the model caused by the scale difference of feature maps (Liu et al., 2019). With the combination of the transformer encoder block, CFFI and ASFF modules, the proposed TIA-YOLOv5 provided the best performance among the compared networks, with the Recall$_{weed}$, F1-score$_{weed}$, AP$_{weed}$ and mAP@0.5 of 90.0%, 70.0%, 80.8% and 90.0%, respectively. Comparing the TIA-YOLOv5 with the baseline network, the improvement of mAP values is not very evident. This is because the mAP is determined by the detection accuracy of both crop and weed, and the AP$_{crop}$ is already 99.2% for the baseline network. With respect to the F1-score$_{weed}$, AP$_{weed}$, the absolute improvement by the TIA-YOLOv5 is 11.8% and 11.3% compared with the baseline network, which is a promising result for weed object detection in the field. In general, the TIA-YOLOv5 network could provide higher weed detection accuracy and remain comparable processing speed relative to the YOLOv5 network.

3.3 Comparison with other object detection networks

The proposed TIA-YOLOv5 was compared with the SSD, Faster RCNN, YOLOv3, YOLOv4 and YOLOv7 networks in terms of Recall$_{weed}$, Precision$_{weed}$, F1-score$_{weed}$, mAP@0.5 and FPS to verify its effectiveness. The results are shown in Table 2. The Faster RCNN is a two-stage object detector that consists of a region of interest (ROI) generation step and a feature extraction step. It has been applied in agriculture for object detection tasks (Li et al., 2021b; Zhao et al., 2022). For the weed and crop detection task in this work, the Faster RCNN yielded an F1-score$_{weed}$ of 60.7%, an mAP of 86.6% and an FPS of 5.8. The detection accuracy is competitive, however, the detection speed is too slow, which should be caused by the separated regression and classification networks of the Faster RCNN (Wang and Liu, 2021), making this model hardly to be used in the field for real-time weed detection. The SSD is a one-stage object detector, and is significantly faster than the Faster RCNN in FPS. However, the SSD did not perform well for small target detection, with the F1-score$_{weed}$ 14% and the mAP 8.6% lower than the TIA-YOLOv5, because it tended to use the low-level feature maps for small target detection. The YOLOv3 (Redmon and Farhadi, 2018) is

TABLE 1 Ablation study on the effect of different modules of the proposed TIA-YOLOv5 network.

Model	Recall$_{weed}$	Precision$_{weed}$	F_1-score$_{weed}$	AP$_{weed}$	mAP@0.5	mAP@0.5:0.95	FPS
YOLOv5	0.757	0.474	0.582	0.695	0.841	0.520	21.5
YOLOv5+Transformer	0.870	0.481	0.619	0.778	0.885	0.579	21.7
YOLOv5+CFFI	0.840	0.480	0.610	0.779	0.885	0.581	19.1
YOLOv5+Transformer+CFFI	0.860	0.492	0.625	0.798	0.894	0.586	20.2
YOLOv5+Transformer+CFFI+ASFF	0.900	0.573	0.700	0.808	0.900	0.580	20.8

TABLE 2 Performance comparison of five networks for sugarbeet and weed detection.

Network	$\text{Recall}_{\text{weed}}$	$\text{Precision}_{\text{weed}}$	$F_1\text{-score}_{\text{weed}}$	mAP@0.5	FPS
Faster RCNN	0.810	0.486	0.607	0.866	5.8
SSD	0.740	0.451	0.560	0.814	12.4
SE-YOLOv3	0.800	0.481	0.600	0.836	24.3
YOLOv4	0.835	0.472	0.603	0.863	21.2
YOLOv7	0.780	0.493	0.604	0.872	19.2
TIA-YOLOv5	0.900	0.573	0.700	0.900	20.8

the most classical network in the YOLO series, and the baseline YOLOv5 in this work was developed from YOLOv3. The YOLOv3 with squeeze-and-excitation networks (SE-YOLOv3) (Hu et al., 2018) uses feature pyramid networks (FPN) to fuse feature maps of different levels. The FPN has a relatively simple architecture and does not need much computing resource, enabling the SE-YOLOv3 has an absolute superiority in detection speed while maintaining acceptable detection accuracy (Wang et al., 2020b). With the appearance of PANet, the combination of the FPN and PANet has significantly improved the model performance on the cross-scale fusion (Wang et al., 2022d). The YOLOv4 network also uses this strategy. Compared with the SE-YOLOv3 and YOLOv4 models, although our proposed TIA-YOLOv5 network performed slightly inferior in FPS, the mAP@0.5 increased by 6.4% and 3.7%, respectively. The newly emerged YOLOv7 network adopts the extended efficient layer aggregation networks (E-ELAN), yielding an $\text{F1-score}_{\text{weed}}$ of 60.4% and an mAP of 87.2%. It was superior to all the compared networks expect the TIA-YOLOv5. With respect to processing speed, when deployed on a Jetson NX (NVIDIA, US) computing platform, the TIA-YOLOv5 model could yield an FPS of about 70 after optimization by the TensorRT SDK (https://github.com/NVIDIA/TensorRT) provided by NVIDIA, which can meet the requirement of real-time application in the field. Overall, our proposed TIA-YOLOv5 provided the best performance for sugarbeet and weed detection among the five compared models.

Several images were selected from the test set for visualizing the detection performance of the TIA-YOLOv5 network, as shown in Figure 8. The yellow boxes in Figures 8 a1 and a2 are the weeds that were missed by the baseline YOLOv5 network but detected by our TIA-YOLOv5. The yellow box in Figure 8 a3 is the region where weed was detected repeatedly by the YOLOv5 while the TIA-YOLOv5 detected them correctly. From this visualization comparison between the YOLOv5 and TIA-YOLOv5, it can be observed that the detection of small object is a challenging task and the YOLOv5 tends to miss these small objects. By adding the transformer encoder block, CFFI and ASFF, the TIA-YOLOv5 has stronger ability for sensing small objects and could effectively avoid problems such as

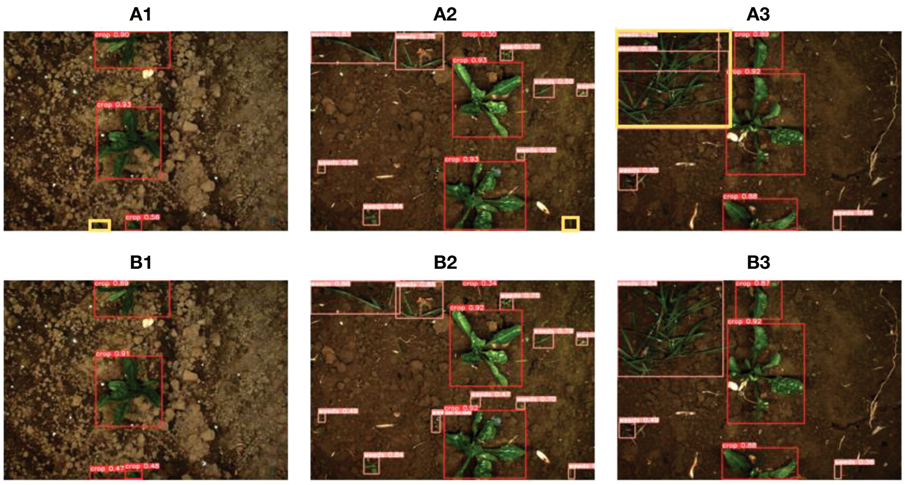

FIGURE 8
Visualization of YOLO5 (Figures a) and TIA-YOLOv5 (Figures b) for detecting sugarbeet and weed.

missed, false and repeated detection, making this model suitable for real-time detection of weeds and crops in the field.

4 Conclusion

In this work, a fast and accurate workflow was proposed for weed and crop object detection in the field. Specifically, a pixel-level synthesization data augmentation method that generated synthetic images by pasting weed pixels into original images was proposed to deal with the problem of unbalanced data distribution of weed and crop. An improved YOLOv5 network named the TIA-YOLOv5 was developed for weed and crop objection detection. In the backbone of the TIA-YOLOv5, a transformer encoder block was used to improve the sensitivity to small weed objects. In the neck, the CFFI was proposed for channel feature fusion and reducing information loss. In the prediction head, the ASFF was introduced for feature fusion of different scales. Test results with a publicly available sugarbeet dataset showed that the proposed TIA-YOLOv5 network yielded an F_1-score$_{weed}$, AP$_{weed}$ and mAP@0.5 of 70.0% 80.8% and 90.0%, respectively, which was 11.8%, 11.3% and 5.9% higher than the baseline YOLOv5 model. And the detection speed reached 20.8 FPS. When deployed on a Jetson NX computing platform, the TIA-YOLOv5 model could yield an FPS of about 70 after optimization by the TensorRT SDK, which is very promising for real-time weed and crop detection in the field. Future work will be focused on developing an SSWM system incorporating the trained weed detection model.

Data availability statement

The data presented in this study are available on request from the corresponding author.

Author contributions

Conceptualization, AW and TP; methodology, AW, TP, and YX; software: TP and YX; validation, AW and TP; formal analysis, AW, TP, and HC; investigation, TP and HC; resources, AW, XW, and BC; data curation, TP and HC; writing—original draft preparation, TP and HC; writing—review and editing, AW; visualization, TP; supervision, AW and YX; project administration, AW and XW; funding acquisition, AW. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Key Research and Development Program of China (No. 2022YFD2000500), National Natural Science Foundation of China (No. 32001417, 61901194), China Postdoctoral Science Foundation (No. 2020M681508, 2022T150276), Jiangsu Postdoctoral Research Funding Program (No. 2021K337C), Jiangsu Modern Agricultural Machinery Equipment and Technology Demonstration and Promotion Program (NJ2021-64), Open Funding from Jiangsu Province and Education Ministry Cosponsored Synergistic Innovation Center of Modern Agricultural Equipment (XTCX2010), and Open Funding from the Key Laboratory of Modern Agricultural Equipment and Technology (Jiangsu University), Ministry of Education (MAET202112, MAET202105).

Acknowledgments

The authors gratefully thank Bin Li for his assistance in training models, Dr. Wenjing Zhu and Dr. Lin Li for their suggestions during the experiment and manuscript preparation.

Conflict of interest

Author YX was employed by company Nanchang Huiyichen Technology Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Angelina, S., Suresh, L. P., and Veni, S. K. (2012). "Image segmentation based on genetic algorithm for region growth and region merging." in 2012 *international conference on computing, electronics and electrical technologies* (Nagercoil:IEEE). 970–974. doi: 10.1109/ICCEET.2012.6203833

Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). YOLOv4: Optimal speed and accuracy of object detection *arXiv preprint. arXiv:2004.10934.* doi: 10.48550/arXiv.2004.10934

Chebrolu, N., Lottes, P., Schaefer, A., Winterhalter, W., Burgard, W., and Stachniss, C. (2017). Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields. *Int. J. Robotics. Res.* 36 (10), 1045–1052. doi: 10.1177/0278364917720510

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al(2020) An image is worth 16x16 words: Transformers for image recognition at scale *arXiv. preprint. arXiv:2010.11929.* doi: 10.48550/arXiv.2010.11929

Gao, J., Liao, W., Nuyttens, D., Lootens, P., Vangeyte, J., Pižurica, A., et al. (2018). Fusion of pixel and object-based features for weed mapping using unmanned aerial vehicle imagery. *Int. J. Appl. Earth observation. geoinformation.* 67, 43–53. doi: 10.1016/j.jag.2017.12.012

Girshick, R. (2015). “Fast r-CNN,” in *Proceedings of the IEEE international conference on computer vision.* (Santiago: IEEE), 1440–1448. doi: 10.1109/ICCV.2015.169

Glenn, J. (2020) *yolov5*. Available at: https://github.com/ultralytics/yolov5.

Hasan, A. M., Sohel, F., Diepeveen, D., Laga, H., and Jones, M. G. (2021). A survey of deep learning techniques for weed detection from images. *Comput. Electron. Agric.* 184, 106067. doi: 10.1016/j.compag.2021.106067

Hu, J., Shen, L., and Sun, G. (2018). “Squeeze-and-Excitation networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition.* (Salt Lake City: IEEE), 7132–7141. doi: 10.1109/CVPR.2018.00745

Jiang, H., Zhang, C., Qiao, Y., Zhang, Z., Zhang, W., and Song, C. (2020). CNN Feature based graph convolutional network for weed and crop recognition in smart farming. *Comput. Electron. Agric.* 174, 105450. doi: 10.1016/j.compag.2020.105450

Kim, Y. H., and Park, K. R. (2022). MTS-CNN: Multi-task semantic segmentation-convolutional neural network for detecting crops and weeds. *Comput. Electron. Agric.* 199, 107146. doi: 10.1016/j.compag.2022.107146

Lee, W.-S., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., and Li, C. (2010). Sensing technologies for precision specialty crop production. *Comput. Electron. Agric.* 74 (1), 2–33. doi: 10.1016/j.compag.2010.08.005

Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., et al. (2021a). “Involution: Inverting the inherence of convolution for visual recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (Nashville: IEEE), 12321–12330. doi: 10.1109/CVPR46437.2021.01214

Li, Z., Li, Y., Yang, Y., Guo, R., Yang, J., Yue, J., et al. (2021b). A high-precision detection method of hydroponic lettuce seedlings status based on improved faster RCNN. *Comput. Electron. Agric.* 182, 106054. doi: 10.1016/j.compag.2021.106054

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*. 2980–2988. doi: 10.1109/TPAMI.2018.2858826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). “SSD: Single shot MultiBox detector,” in *European Conference on computer vision* (Cham: Springer), 21–37. doi: 10.1007/978-3-319-46448-0_2

Liu, S., Huang, D., and Wang, Y. (2019) Learning spatial fusion for single-shot object detection *arXiv. preprint. arXiv:1911.09516.* doi: 10.48550/arXiv.1911.09516

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). “Path aggregation network for instance segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8759–8768.

Lottes, P., Behley, J., Milioto, A., and Stachniss, C. (2018). Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics. Automation. Lett.* 3 (4), 2870–2877. doi: 10.1109/LRA.2018.2846289

Luo, Y., Cao, X., Zhang, J., Guo, J., Shen, H., Wang, T., et al. (2022). CE-FPN: enhancing channel information for object detection. *Multimedia. Tools Appl.* 81 (*21*), 1–20. doi: 10.1007/s11042-022-11940-1

McCarthy, C., Rees, S., and Baillie, C. (2010). “Machine vision-based weed spot spraying: a review and where next for sugarcane?,” in *Proceedings of the 32nd Annual Conference of the Australian Society of Sugar Cane Technologists (ASSCT 2010)*. (Bundaberg, Australia), 424–432.

Peng, H., Li, Z., Zhou, Z., and Shao, Y. (2022). Weed detection in paddy field using an improved RetinaNet network. *Comput. Electron. Agric.* 199. doi: 10.1016/j.compag.2022.107179

Redmon, J., and Farhadi, A. (2018) YOLOv3: An incremental improvement *arXiv. preprint. arXiv:1804.02767.* doi: 10.48550/arXiv.1804.02767

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). “Generalized intersection over union: A metric and a loss for bounding box regression,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (Long Beach: IEEE), 658–666. doi: 10.1109/CVPR.2019.00075

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2022a) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors *arXiv. preprint. arXiv:2207.02696.* doi: 10.48550/arXiv.2207.02696

Wang, Q., Cheng, M., Huang, S., Cai, Z., Zhang, J., and Yuan, H. (2022b). A deep learning approach incorporating YOLOv5 and attention mechanisms for field real-time detection of the invasive weed solanum rostratum dunal seedlings. *Comput. Electron. Agric.* 199, 107194. doi: 10.1016/j.compag.2022.107194

Wang, A., Li, W., Men, X., Gao, B., Xu, Y., and Wei, X. (2022c). Vegetation detection based on spectral information and development of a low-cost vegetation sensor for selective spraying *Pest Manage. Science.* 78, 2467–2476. doi: 10.1002/ps.6874

Wang, X., and Liu, J. (2021). Tomato anomalies detection in greenhouse scenarios based on YOLO-dense. *Front. Plant Sci.* 12, 634103. doi: 10.3389/fpls.2021.634103

Wang, A., Xu, Y., Wei, X., and Cui, B. (2020a). Semantic segmentation of crop and weed using an encoder-decoder network and image enhancement method under uncontrolled outdoor illumination. *IEEE Access* 8, 81724–81734. doi: 10.1109/access.2020.2991354

Wang, A., Zhang, W., and Wei, X. (2019). A review on weed detection using ground-based machine vision and image processing techniques. *Comput. Electron. Agric.* 158, 226–240. doi: 10.1016/j.compag.2019.02.005

Wang, Y., Zhang, Z., and Yin, H. (2020b). Detection method of dense bridge disease targets based on SE-YOLOv3. *J. Phys.* 1544 (1), 012141. doi: 10.1088/1742-6596/1544/1/012141

Wang, H., Zhang, S., Zhao, S., Wang, Q., Li, D., and Zhao, R. (2022d). Real-time detection and tracking of fish abnormal behavior based on improved YOLOV5 and SiamRPN++. *Comput. Electron. Agric.* 192, 106512. doi: 10.1016/j.compag.2021.106512

Wang, X., Zhao, Q., Jiang, P., Zheng, Y., Yuan, L., and Yuan, P. (2022e). LDS-YOLO: A lightweight small object detection method for dead trees from shelter forest. *Comput. Electron. Agric.* 198, 107035. doi: 10.1016/j.compag.2022.107035

Zhang, Z. P. (2003). Development of chemical weed control and integrated weed management in China. *Weed. Biol. Manage.* 3 (4), 197–203. doi: 10.1046/j.1444-6162.2003.00105.x

Zhao, S., Liu, J., and Wu, S. (2022). Multiple disease detection method for greenhouse-cultivated strawberry based on multiscale feature fusion faster R_CNN. *Comput. Electron. Agric.* 199, 107176. doi: 10.1016/j.compag.2022.107176

Zhu, X., Lyu, S., Wang, X., and Zhao, Q. (2021). “TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (Monreal: IEEE), 2778–2788. doi: 10.1109/ICCVW54120.2021.00312