IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

A Robust Method for Wheatear Detection Using UAV in Natural Scenes

MING-XIANG HE[1,2*], PENG HAO[1,*], YOU-ZHI XIN[1]

[1]College of Computer Science and Engineering, Shandong University of Science and Technology, Tsingtao 266590, China
[2]National Virtual Simulation Experiment Center of Shandong University of Science and Technology, Shandong University of Science and Technology, Tsingtao 266590, China

Corresponding author: YOU-ZHI XIN (xyz3816@163.com).

Author Contributions: *MING-XIANG HE, PENG HAO contributed equally.

This work was supported in part by the Industry-University Cooperation Education Program of the Ministry of Education under Grant 201901035016 and the 2020 Qingdao social science planning research project under Grant QDSKL2001143.

ABSTRACT In recent years, deep learning has greatly improved the ability of wheatear detection. However, there are still three main problems in wheatear detection based on unmanned aerial vehicle (UAV) platforms. First, dense wheat plants often overlap, and the wind direction will blur the pictures, which obviously interferes with the detection of wheatears; second, due to the different maturity, color, genotype, and head orientation, the appearance will also be different; third, UAV needs to take images in the field and conduct real-time detection, which requires the embedded module to detect wheatears quickly and accurately. Given the above problems, we studied and improved YoloV4, and proposed a robust method for wheatear detection using UAV in natural scenes. For the first problem, we modified the network structure, deleted the feature map with a size of 19×19, and used k-means algorithm to re-cluster the anchors, and we proposed a method of prediction box fusion. For the second problem, we used the pseudo-labeling method and data augmentation methods to improve the generalization ability of the model. For the third problem, we simplified the network structure, replaced the original network convolution with the improved depthwise separable convolution, and proposed an adaptive ReLU activation function to reduce the amount of calculation and speed up the calculation. The experimental results showed that our method can effectively mark the bounding of wheatears. In test sets, our method achieves 96.71% in f1-score, which is 9.61% higher than the state of the art method, and the detection speed is 23% faster than the original method. It can be concluded that our method can effectively solve the problems of wheatear detection based on the UAV platform in natural scenes.

INDEX TERMS Wheatear detection, improved YoloV4, UAV, object detection, deep learning

I. INTRODUCTION

Wheat is a kind of cereal crop widely planted all over the world, and it is one of the staple foods of human beings. It is particularly important to effectively monitor the growth of wheat and make scientific management decisions in a modern society of rapid population growth. The manual detection method is time-consuming, laborious, the observation area is small, and the accuracy is biased; moreover, with the increase of planting scale, it becomes more and more infeasible. UAV single sorties operation time is long, the coverage is very wide, with high efficiency and flexibility, easy to operate, and other advantages, so it has been widely used in the agricultural field [1]-[3]. In order to effectively analyze the growth of wheat, the wheatear detection equipment based on UAV platform can collect wheat images in the field and detect the wheatears in the images in real-time. With these detection results, farmers can estimate wheat density and the size of wheatears, assess wheat health and maturity, and make appropriate management decisions accordingly. Therefore, it is of great significance and application value for research on a wheatear detection model which is suitable for UAV platform with good performance.

The accuracy and speed of wheatear detection are the primary tasks and the main design difficulty of wheatear detection model, and they are directly related to the work efficiency of wheatear detection UAV. At present, crop detection using UAV is mainly based on spectral remote sensing technology [4]-[6]. Although it can detect the overall situation of crops within a certain scope, it cannot accurately

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

detect each wheat plant, and there are great limitations in light environment and angle when shooting. Our goal is to accurately detect wheatears, to more accurately judge the wheat growth and health status. With the development of science and technology, more and more object detection methods are being proposed. Due to different sunlight exposure, maturity, genotype and head orientation, the color, and shape of wheatears will differ, and there will be overlapping of wheatears caused by dense plants. All these will affect the accuracy of wheatear detection. In practical applications, due to hardware limitations, it is often necessary to sacrifice accuracy to ensure the inference speed of the detector. Therefore, the balance between the effectiveness and efficiency of the target detector must be considered.

YoloV4 [7] integrates various advanced technologies based on YoloV3 [8] and provides the best trade-off in terms of speed and accuracy. To solve the previous problems, we have conducted in-depth research on the various sub-modules of yolov4, so that the improved method can accurately mark the wheatears in the images taken by the UAV in the field, and its performance is better than that of YoloV4. We proposed a robust wheatear detection method using UAV in natural scenes. The main contributions of this paper include the following four points:

1) We used improved depthwise separable convolution to replace the original network convolution, to reduce the number of parameters and speed up the calculation.

2) Based on YoloV4, the network architecture was adjusted. According to the characteristics of wheatear objects, the feature map with a resolution size of 19×19 was deleted, and the k-means algorithm was adopted to cluster the bounding boxes to replace the original anchors, which simplify the network structure and improves the performance.

3) We proposed an adaptive ReLU activation function, which determines the appropriate activation function according to all input elements. It does not increase the depth and width of the network, and is more accurate than ReLU and faster than Mish.

4) We proposed a method of prediction box fusion, which was adopted to solve the problem that all predicted bounding boxes are inaccurate in the test stage, and improved the detection accuracy under overlapping condition.

The rest of this paper is structured as follows: in chapter two, we reviewed the work related to our research content; in chapter three, we introduced in detail the improved methods in this paper; in chapter four, we introduced the data sets, evaluation indicators, and tricks used in the experiment, and conducted the ablation study; in chapter five, we designed a large number of contrast experiments and analyzed the experimental results; in chapter six, we summarized the research work of this paper.

II. RELATED WORK

A. DETECTION OF CROPS

Wheatear detection is an important means to monitor wheat growth and health. The main methods are artificial field detection and automatic wheatear detection based on images. Due to the time-consuming, laborious, and subjectivity of artificial detection methods, they are not popular in today's rapid development of agriculture. Image-based detection technology has been widely recognized in crop detection in recent years due to its strong real-time performance and robustness [9]-[12]. Generally, there are two methods for image-based object detection: the traditional manual feature-based object detection method [13]-[14] and the deep learning-based object detection method [15]-[17].

Traditional object detection methods are generally divided into three steps: region selection, feature extraction of candidate region, and classifier classification [18]. For example, Li et al. combined the texture features of wheatears with neural networks to detect wheatears in a laboratory environment with whiteboard background [19], while Zhou et al. combined multi-feature optimization and twin support vector machine to detect wheatears [20]. This kind of method can detect the wheatears in the image, but it uses manual features to represent the object features, so it has poor robustness. The different colors and shapes of wheatears will cause serious deviation in the results of these methods. Moreover, the method based on the sliding window region selection strategy will lead to many redundant calculations, and the whole process has high complexity.

With the development of deep learning, a variety of object detection methods have been proposed. The object detection method based on deep learning can extract complex feature hierarchies from images through self-learning [21], which can effectively solve the problem of the poor performance of manual feature extraction methods and has been widely used in agricultural crop growth and health monitoring. Object detection methods based on deep learning can be divided into two categories: two-stage object detection method and one-stage object detection method. The two-stage object detection method divides the object detection into two stages, that is, the region proposal network (RPN) is used to extract the candidate object information, and then the detection network is used to predict and identify the location and category of the candidate object. For example, Lootens used machine learning based on UAV images to automatically count wheatears [22], but this method cannot accurately mark the wheatears in the image. Ma et al. combine DCNN and FCN to segment wheat ears [23], this method is divided into two stages, the first stage uses DCNN for coarse segmentation, and the second stage uses FCN for fine segmentation, its accuracy is good, but the speed is too slow. Madec et al. proposed an ear density estimating method by using Faster-RCNN and RGB images of high spatial resolution [24], Wang et al. combined FCN and Harris Corner Detection to detect wheatears in the field [25], but

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

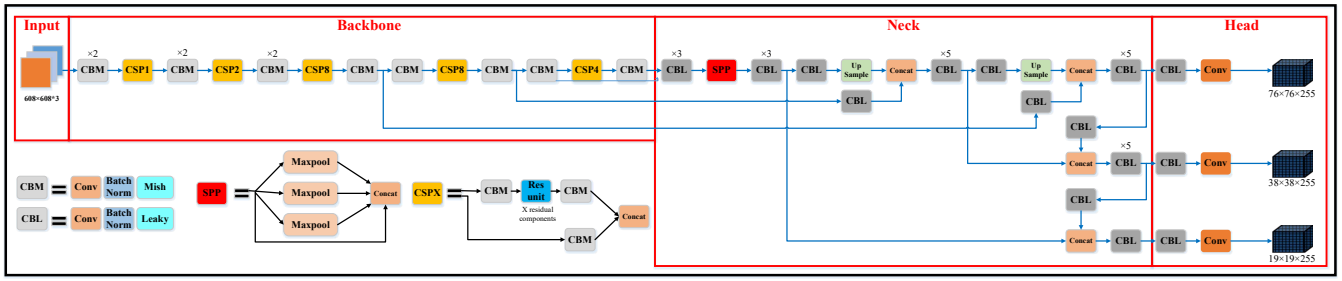


FIGURE 1. The overall network structure of YoloV4.

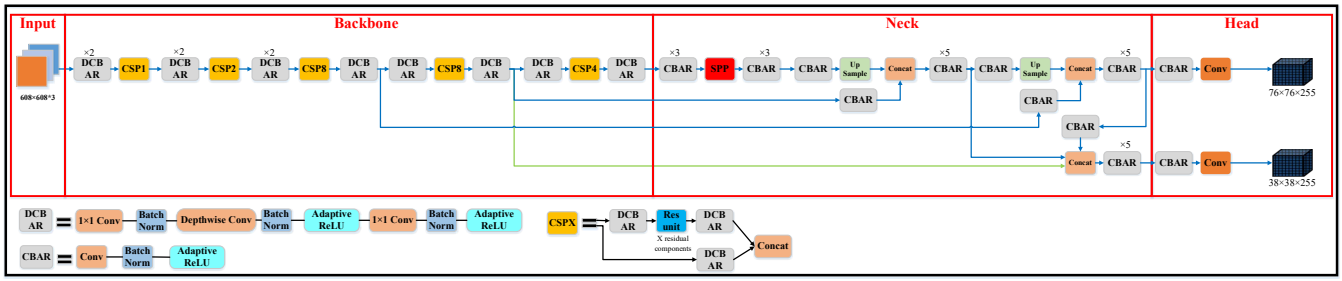


FIGURE 2. The overall network structure of our method.

these two methods were focusing on counting the number of wheatears. Although they have high detection accuracy and strong generalization, due to there are many regional suggestions, high requirements for equipment performance, and low detection efficiency, they cannot meet the real-time requirements. The one-stage object detection method uses the regression method based on deep learning to detect the object and obtains the boundary box and object type directly from the image, which greatly improves the speed of object detection. For example, Yang et al. carried out field wheatear detection based on YoloV3 [26]. This kind of method has higher detection accuracy, stronger generalization ability, and higher detection speed. However, for the detection of wheatears, in the outdoor field, wheatears are often covered or overlapped, In this case, the accuracy of these methods is greatly reduced. If it is applied to the wheatear detection UAV, the detection results will mislead the wheat managers and make wrong decisions.

B. INTRODUCTION TO YOLOV4

YoloV4 is the fourth iteration of Yolo [27]. It improves the image pyramid to feature the pyramid and integrates various advanced algorithms. Compared with YoloV3, YoloV4 has better detection performance for occluded and overlapped objects. Since the main difficulty of wheatear detection is the overlapping and occlusion of wheatears, YoloV4 is more suitable as the wheatear detection module embedded in wheatear detection UAV than YoloV3. The network structure of YoloV4 is shown in Figure 1. Next, we will introduce the Backbone, Neck, and Head modules of YoloV4.

1) The Backbone module is mainly used to extract rich information features from the input image. The Backbone of YoloV4 is based on the YoloV3 backbone network and the experience of Cross Stage Partial Networks (CSPNet) [28]. CSPNet solves the problem of repeated gradient information in the Backbone of other large-scale convolutional neural network framework and integrates the gradient changes into the feature map from the beginning to the end. Therefore, the parameters and FLOPS values of the module are reduced, which not only ensures the reasoning speed and accuracy but also reduces the module size.

2) The Neck module is mainly used to generate feature pyramids to enhance the module's detection of objects with different scales to identify the same object with different sizes and scales. Based on the idea of Path Aggregation Network for Instance Segmentation (PANET) [29], the Neck of YoloV4 enhanced information dissemination based on the framework of Mask R-CNN [30] and FPN [31].

3) The Head module is mainly used in the final detection part. YoloV4 uses the Head of YoloV3 in the Head module. It applies anchor on the feature map and generates the final output vector with class probabilities, object values, and bounding box. These feature vectors with different scaling scales are used to detect objects of different sizes.

III. PROPOSED METHOD

In this chapter, we will describe our improvement points in detail, and illustrate their advantages by comparing them with the methods in the original network, and we will conduct the ablation study to verify the rationality of these improvements in chapter four. The network structure of our method is shown in Figure 2.

A. BACKBONE NETWORK

Convolution layer is an important part of the whole neural network. It can automatically extract complex feature

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

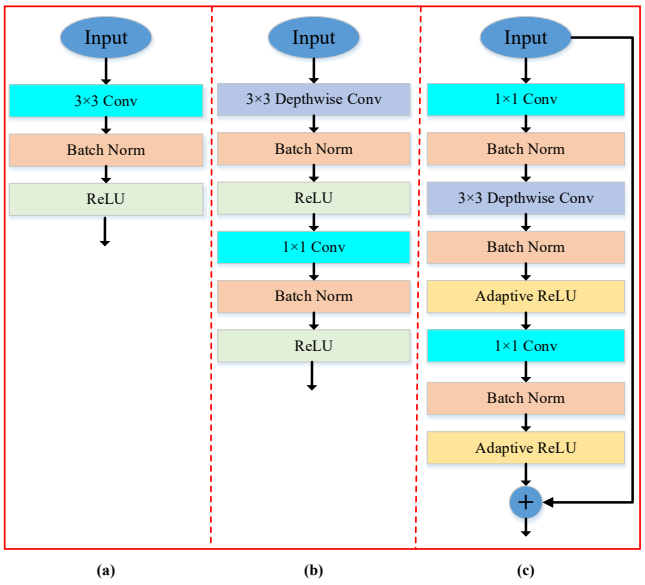


FIGURE 3. (a) Original standard convolution, (b) Depthwise separable convolution, (c) Improved depthwise separable convolution.

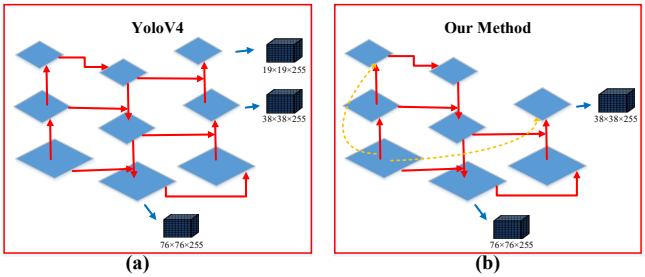


FIGURE 4. Comparison between the original method (a) and the improved (b) network structure.

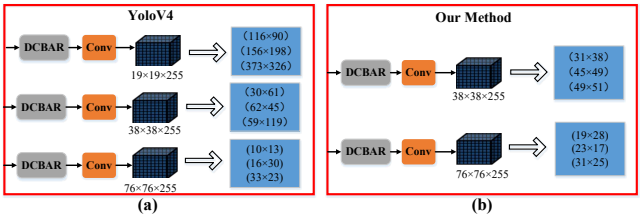


FIGURE 5. The original method (a) and modified (b) anchor comparison for different resolution feature maps.

hierarchies from images. However, there are too many parameters, and the calculation is complex in traditional convolution. To speed up the calculation and save the calculation cost, we proposed an improved depthwise separable convolution to replace the traditional convolution. The structure comparison is shown in Figure 3.

In Figure 3, we set the number of input channels as 3 and the number of output channels as 256. The traditional convolution is directly connected with a $3 \times 3 \times 256$ convolution kernel, parameters are $3 \times 3 \times 3 \times 256 = 6912$; The improved depthwise separable convolution is completed in three steps: 1×1 Convolution process, 3×3 depthwise convolution process, and 1×1 convolution process. The number of parameters is $1 \times 1 \times 1 \times 256 + 3 \times 3 \times 3 + 3 \times 1 \times 1 \times 256 = 1051$, which is much less than that of the traditional method. This method greatly improves computing efficiency. In addition, the 1×1 convolution is added to increase the depth of the network and add nonlinearity without increasing the receptive field.

B. IMPROVED FEATURE FUSION STRUCTURE

YoloV4 uses the structure of feature pyramid network PANET to fuse the deep level feature information with the shallow feature information and uses the multi-scale fusion method to predict the location and category on the multi-scale feature map. However, the three-scale feature fusion method adopted by YoloV4 network structure has adverse effects on small objects such as wheatears, the semantic loss of feature map with a resolution size of 19×19 is serious, which easily leads to the loss of small objects. Considering that the resolution of the feature map will directly affect the small object detection and the overall performance, we modify the resolution of three scales of the feature map from $19 \times 19, 38 \times 38, 76 \times 76$ to $38 \times 38, 76 \times 76$. Base on the original network structure, we have added two channels, as shown by the orange dashed line in Figure 4, which can fuse more features without increasing the computational cost. The modified network structure is shown in Figure 4.

The feature maps of YoloV4 with different sizes have three anchors respectively, and larger feature maps use smaller anchors to get more edge information of the object. However, the anchors defined by the original YoloV4 network and the hierarchical structure of the network cannot be well applied to the research object of this paper. So, we used the k-means algorithm to cluster the wheatear dataset and replace the original network anchors with the clustered anchors to improve the accuracy of the predicted bounding boxes. Based on the modification of the above network, we set the k-means clustering algorithm with the clustering category as 6, and new anchors obtained are $(19 \times 28), (23 \times 17), (31 \times 25), (31 \times 38), (45 \times 49), (59 \times 51)$, the anchor structure comparison of different resolution feature maps after clustering is shown in Figure 5.

C. ADAPTIVE RELU ACTIVATION FUNCTION

The selection of activation functions is critical for deep learning network. Based on ReLU, we proposed an adaptive activation function, which determines the appropriate activation function according to all input elements. Compared with Mish, it has fewer calculations, and compared with ReLU, it has higher accuracy. The calculation of the ReLU function is shown in Equation (1).

$$ReLU = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{1}$$

And it can be generalized as Equation (2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

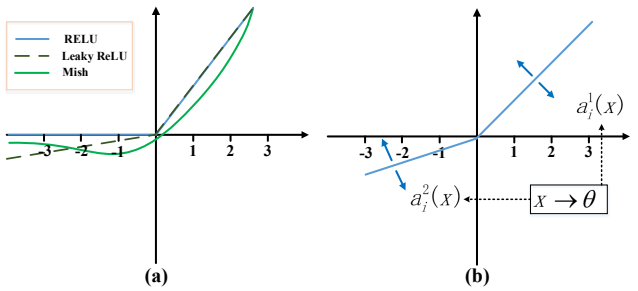


FIGURE 6. The function used for comparison is (a), the adaptive ReLU function is (b).

$$y_i = \max\{a_i^j \chi_i + b_i^j\} \quad (2)$$

Where x is the input vector, for the input vector x_i of the i[th] channel, the activation is calculated as $y_i = \max\{x_i, 0\}$. Based on Equation (2), we propose an adaptive ReLU activation function, which is a further extension of ReLU, its calculation is shown in Equation (3).

$$y_i = f_{\theta(x)}(x_i) = \max\{a_i^j(\boldsymbol{x})x_i + b_i^j(\boldsymbol{x})\},\ 1 \le k \le K \quad (3)$$

Where $\boldsymbol{x} = \{x_i\}$ are all input elements, i is the number of channels, and j is the number of functions, function $\boldsymbol{\theta}(\boldsymbol{x}) = [a_1^1, \dots, a_i^1, \dots, a_1^j, \dots, a_i^j, \dots, b_1^1, \dots, b_i^1, \dots, b_1^j, \dots, b_i^j]^T$, a_i^j and b_i^j are the output of function $\boldsymbol{\theta}(\boldsymbol{x})$ and are the sum of initialization and residual, and its calculation is shown in Equation (4).

$$a_i^j(\boldsymbol{x}) = \alpha^j + \gamma_a \Delta a_i^j(\boldsymbol{x}),\ b_i^j(\boldsymbol{x}) = \beta^j + \gamma_b \Delta b_i^j(\boldsymbol{x}) \quad (4)$$

Where α^j and β^j are the initial values of a_i^j, γ_a and γ_b are the scalars controlling the residual range. For example, if $j = 2, \alpha^1 = 1, \alpha^2 = \beta^1 = \beta^2 = 0$, corresponding to ReLU, the default γ_a and γ_b are 1 and 0.5, respectively.

The adaptive ReLU does not increase the depth and width of the network, but can effectively improve the performance of the model. Through the test, compared with ReLU, the accuracy is improved by 4.3%, and the speed is increased by 3.4% compared with mish. The function comparison image is shown in Figure 6.

D. METHOD OF PREDICTION BOX FUSION

During the testing phase, when we input the image to be detected into the model, we get a set of predicted bounding boxes. For images containing many wheatears, the predicted bounding boxes produced on a single wheatear object may not be accurate due to the overlap of wheatears. To solve this problem, we propose a method of prediction bounding box fusion. This method uses information from all predicted bounding boxes to solve the problem of inaccuracy, as shown in Figure 7. We set the confidence of a fusion box to the average confidence of all the boxes that constitute it. The calculation is shown in Equation (5).The coordinates of the fusion box are the weighted sum of the coordinates of each

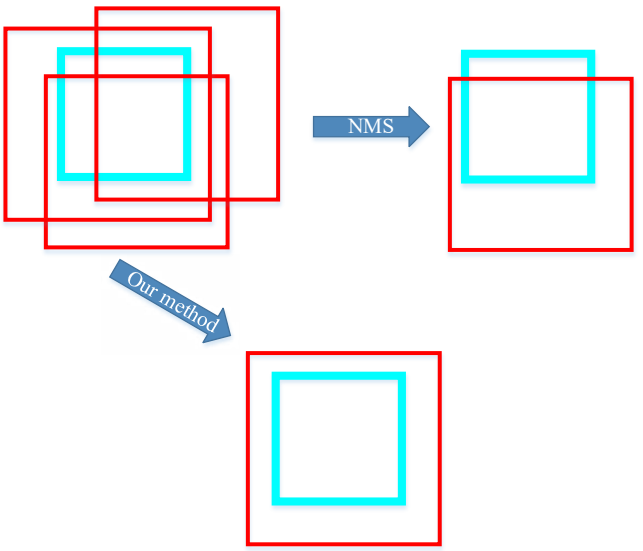


FIGURE 7. Defining the predicted bounding boxes using method of prediction box fusion.

prediction bounding box, where the weight is the confidence of the box. The calculation is shown in Equation (6) and (7). Boxes with high confidence must have a greater impact on the fusion box coordinates, not boxes with low confidence. By testing on this test set, this method improved the accuracy of the bounding box by 8.3% compared with the NMS method.

$$C = \frac{C_1 + C_2 + \cdots + C_n}{n} \quad (5)$$

$$\boldsymbol{X_1} = \frac{C_1 X_{1_1} + C_2 X_{1_2} + \cdots + C_n X_{1_n}}{C_1 + C_2 + \ldots + C_n},\ \boldsymbol{X_2} = \frac{C_1 X_{2_1} + C_2 X_{2_2} + \cdots + C_n X_{2_n}}{C_1 + C_2 + \ldots + C_n} \quad (6)$$

$$\boldsymbol{Y_1} = \frac{C_1 Y_{1_1} + C_2 Y_{1_2} + \cdots + C_n Y_{1_n}}{C_1 + C_2 + \ldots + C_n},\ \boldsymbol{Y_2} = \frac{C_1 Y_{2_1} + C_2 Y_{2_2} + \cdots + C_n Y_{2_n}}{C_1 + C_2 + \ldots + C_n} \quad (7)$$

Where n is the number of prediction boxes before fusion, C_n is the confidence of each prediction bounding box, $\boldsymbol{X_1}, \boldsymbol{X_2}$, $\boldsymbol{Y_1}, \boldsymbol{Y_2}$ are the coordinate values of the fusion box, and $C_n X_{1_n}$, $C_n X_{2_n}$, $C_n Y_{1_n}$, $C_n Y_{2_n}$ are the product of the coordinate value of each corresponding prediction box and the confidence of each corresponding box.

IV. EXPERIMENTS

We give a flowchart of this method based on the above improvements, as shown in Figure 8. In this experiment, based on Intel i7-8750H CPU and NVIDIA GTX 1080 GPU, we build a Darknet deep learning framework with Python 3.7 development environment under Linux operating system. The training and detection program of wheatear detection network model based on YoloV4 was written in Python language.

A. DATA SET ANALYSIS

Our experiment uses the Global Wheat Head Detection (GWHD) [32] as the data set. It contains 4,700 high-resolution RGB images and 190,000 labelled wheatears collected from several countries around the world at different growth stages with a wide range of genotypes. We use 3300 images as the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

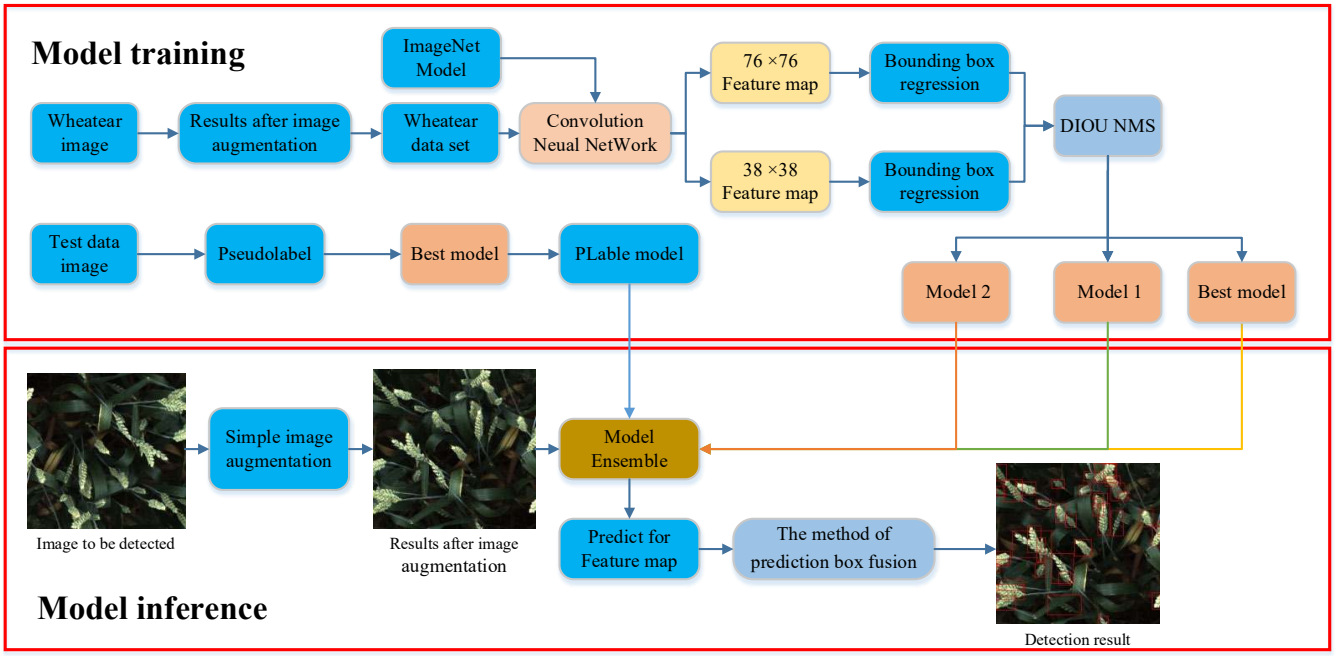


FIGURE 8. **The training and detecting process of the wheatear detection method.**

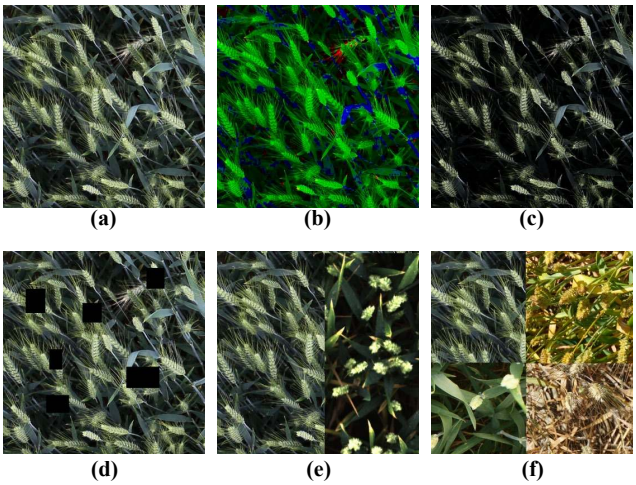


FIGURE 9. **(a) Original image, (b) Image after using HSV, (c) Image after brightness conversion, (d) Image after Cutout, (e) Image for Cutmix, (f) image for Mosaic.**

training set for the experiment and 1400 images as the test set.

We analyzed the samples in the data set. Due to the different shooting environment, some samples have bright colors and some are dim; due to different maturity periods and types, the color and appearance of wheatears are different, some have wheat awns, some do not, some are green, and some are yellow; due to the different shooting height and angle, the number of wheatears in the image is different.

B. DATA AUGMENTATION

In deep learning, enough training samples are usually required. The larger the number of samples, the better the trained model and the stronger the generalization ability. However, in actual training, training samples are always limited, and it is impossible to capture an image for every real-world scene. Therefore, it is necessary to augment the existing training data to generalize to other situations, thereby allowing the model to adapt to a wider range of situations. We used some methods to expand the training samples. The effect is shown in Figure 9. The specific methods are as follows:

1) HSV channel color conversion.

2) Brightness and contrast conversion.

3) Horizontal flip, vertical flip, grayscale conversion, and random cropping.

4) Using the Cutout [33] method, randomly cut some areas of the sample and fill them with 0 pixels, and the result of the classification remains unchanged; use the Cutmix [34] method to cut some areas without filling 0 pixels, but randomly fill in the regional pixel values of other data in the training data set, and the classification results are distributed in a certain proportion, which can effectively prevent overfitting.

5) Using the Mosaic method [7], four training images are combined into one image according to a certain ratio to enrich the background of the detected object, so that the model can learn to recognize the object in a smaller range. In addition, batch normalization calculates activation statistics from 4 different images on each layer. This significantly reduces the need for a large mini-batch size, it can use one GPU to achieve better results.

C. SELECTION OF TRICKS

This article used several tricks to improve the performance of our model, as shown below:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

Larger Batch Size Using a larger batch size can improve the stability of training and get better results. Here we change the training batch size from 64 to 192, and adjust the training schedule and learning rate accordingly.

EMA When training a model, it is often beneficial to maintain moving averages of the trained parameters. Evaluations that use averaged parameters sometimes produce significantly better results than the final trained values [35]. The Exponential Moving Average (EMA) compute the moving averages of trained parameters using exponential decay. The calculation for each parameter is shown in Equation (8).

$$W_{EMA} = \lambda W_{EMA} + (1-\lambda)W \tag{8}$$

where λ is the decay. We apply EMA with decay λ of 0.9998 and use the shadow parameter W_{EMA} for evaluation.

Pseudo-Labelling Based on the optimized model getting during the training stage, we used the Pseudo-Labelling [36] method on the test set to further optimize the model to improve its generalization ability under complex test data. Pseudo-Labelling is defined by Semi-supervised learning. The core idea of Semi-supervised learning is to improve the generalization ability of the model in the supervised process by using labeled data. Pseudo-Labelling is a process that uses a trained model to make predictions on unlabeled data, and samples are screened based on the predicted results and re-input into the model for training.

D. EVALUATION INDICATORS

In this experiment, we used f1-score and the average precision of a single image (Avg-P) evaluation indexes to evaluate our method and other comparison methods.

The calculation formula of f1-score(F1) is as Equation (9).

$$P = \frac{TP(t)}{TP(t)+FP(t)},\ R = \frac{TP(t)}{TP(t)+FN(t)},\ F1 = \frac{2PR}{P+R} \tag{9}$$

In formula (9), P is the accuracy rate, R is the recall rate, TP is the number of true positive samples, FP is the number of false positive samples, and FN is the number of false negative samples. When a single predicted object matches a ground-truth object whose IOU is higher than the threshold, a true positive is calculated. False positive means that the predicted object has no related ground truth objects. False negative means the basic real object without relevant prediction. The IOU computational formulas for predicted bounding boxes and ground truth bounding boxes is shown as Equation (10).

$$IOU(A,B) = \frac{A \cap B}{A \cup B} \tag{10}$$

The calculation method of Avg-P is: Scan a series of IOU thresholds by measurement and calculate an average accuracy value at each point. The threshold range is 0.5 to 0.75, and the step size is 0.05. In other words, when the threshold is 0.5, if the intersection of the predicted object and the ground truth object is greater than 0.5, the predicted object is regarded as a "hit". If there are no real objects on the ground at all in a given image, any number of predictions (false positives) will result in an image score of zero and be included in the average accuracy. This indicator can not only evaluate whether the detection result is accurate, but also the accuracy of the final output bounding box. The calculation method is as Equation (11).

$$Avg\text{-}P = \frac{1}{|thresholds|}\sum_{t}\frac{TP(t)}{TP(t)+FP(t)+FN(t)} \tag{11}$$

TABLE 1. The results of the ablation study.

	Methods	F1%	Avg-p%	FPS
A	YOLOV4	88.23	62.75	57
B	A + Improved depthwise separable convolution	88.19	63.21	64
C	B + Modified network and anchors re-clustered by k-means	92.15	69.19	70
D	D + Adaptive ReLU	93.31	70.05	72
E	E + method of prediction box fusion	96.71	77.81	72

E. ABLATION STUDY

In this part, we present the effectiveness of each module in an incremental manner. The reason is that each improved method is not completely independent. Some improved methods are effective when applied alone, but they are not effective when combined together. Since there are too many combinations of various improved methods, it is difficult to conduct a comprehensive analysis. Therefore, we carry out an ablation study to prove the rationality of our method combination, the results are shown in Table 1.

A → B First of all, we replace the original YoloV4 traditional convolution with the improved depthwise separable convolution. We found that compared with the traditional convolution, although the accuracy is not improved, the calculation speed is faster.

B → C We try to modify the network structure and delete the 19×19 feature map to make the modified network structure simpler, and used anchors clustered by K-means algorithm , to make it more suitable for the characteristics of the small wheatear target. We can see that f1-score, avg-p and FPS have all been significantly improved.

C → D We use adaptive ReLU to replace the original activation function, which not only does not increase the depth and width of the network, but also can effectively improve the performance of the model. We found that accuracy and speed have been improved.

E → F After using the method of prediction box fusion method, we found from the avg-p that the accuracy has been greatly improved. It shows that this method can greatly improve the problem which all prediction boxes are inaccurate when wheat ears are dense.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

This study shows that the combination of our six improvement methods has greatly improved the accuracy and speed compared with the original method.

V. ANALYSIS OF EXPERIMENTAL RESULTS

In this section, we compared our method with previous related methods (Method of original YoloV4 [7], method of Zhu et al. [10], method of Madec et al. [24], and method of Yang et al. [26]). The models of these methods were all obtained in the same environment using the same training strategy. We used the number of wheatears in a single image, different lighting environments and wheat maturity periods were respectively as control variables to verify the performance of our method in wheatear detection in the field.

A. COMPARISON OF DETECTION RESULTS UNDER DIFFERENT NUMBERS OF WHEATEARS

In the actual detection process of the wheatear detection UAV, due to the shooting angle, distance, and other factors, the number of wheatears in the pictures taken are different. When the number of wheatears is small and the volume is large, the detection object is clear, complete, and less overlapping, which is convenient for detection. However, in actual situations, due to the increase in the number of objects and the decrease in size, overlap and occlusion may occur, making detection difficult. To this end, we established contrast experiments under different numbers of wheatears, which are: wheatear detection under sparse numbers, wheatear detection under normal numbers, and wheatear detection under dense numbers, respectively. We compared the wheatear detection performance of various methods under different numbers of wheatears.

In the test set of this experiment, we selected 360 pictures of the same wheat species and a similar light environment, containing 18,732 wheatears, and divided them into three groups according to the number of wheatears in a single image. There were 120 pictures with the number of wheatears less than 10, 120 pictures with the number of wheatears between 30 and 60, and 120 pictures with more than 80 wheatears.

We carried out three experiments with several different methods, each time taking 75 pictures randomly selected from each test set as the experimental test set. We calculated the parameters rate, and the recall rate to get the f1-Score in each experiment. Finally, we averaged three types of results to obtain a comprehensive result, as shown in Table 2.

As can be seen from Table 2, our method shows superior performance. When the number of wheatears is sparse, all methods have good performance. When the number of wheatears increases from less than 10 to more than 40, the method of Zhu et al. [10] performed the worst. The f1-score of our method was relatively stable, and the methods of other people had a significant decline. However, when the number of wheatears increased to more than 80, the f1-score of our method dropped by 5%, while the others' methods dropped by about 10%. Because the wheat is dense in the image and there are many occlusions, there are many wheatears that are not detected. Through a comprehensive analysis of the results, we can conclude that our method is more suitable for different wheatear numbers. The test results are shown in Figure 10.

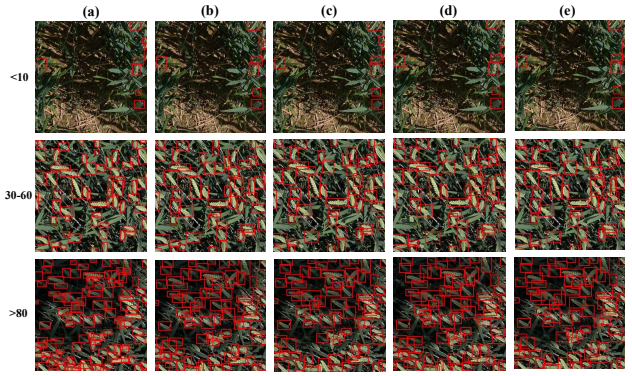


FIGURE 10. Detection results of different methods when the numbers of wheatears are different: (a) our method, (b) YoloV4, (c) method of Zhu et al., (d) method of Madec et al., (e) method of Yang et al..

TABLE 2. Detection results of different methods with different numbers of wheatears.

Number	Method	F1%			
		1	2	3	Average
<10	Our Method	98.47	98.53	98.46	**98.49**
	YoloV4 [7]	96.33	96.37	96.42	96.37
	Zhu et al. [10]	93.62	93.55	93.49	93.55
	Madec et al. [24]	96.27	96.19	96.22	96.23
	Yang et al. [26]	95.31	95.36	95.34	95.34
30-60	Our Method	96.73	96.76	96.69	**96.73**
	YoloV4 [7]	90.57	90.53	90.55	90.55
	Zhu et al. [10]	83.31	83.38	83.37	83.35
	Madec et al. [24]	91.17	91.16	91.17	91.17
	Yang et al. [26]	88.23	88.19	88.21	88.21
>80	Our Method	91.35	91.42	91.37	**91.38**
	YoloV4 [7]	88.63	88.61	88.63	88.62
	Zhu et al. [10]	76.12	76.15	76.13	76.13
	Madec et al. [24]	82.31	82.35	82.32	82.33
	Yang et al. [26]	78.27	78.31	78.29	78.29
Average	Our Method	95.52	95.57	95.51	**95.53**
	YoloV4 [7]	89.51	89.50	89.53	89.52
	Zhu et al. [10]	84.35	84.36	84.33	84.35
	Madec et al. [24]	89.92	89.90	89.90	89.91
	Yang et al. [26]	87.27	87.29	87.28	87.28

B. COMPARISON OF DETECTION RESULTS IN DIFFERENT LIGHT ENVIRONMENTS

In the experiment of this chapter, we used different lighting environments as control variables, with the light varies from bright to dim due to different shooting conditions. Under normal light conditions, the wheatears are clearly visible, and the detection is simple. However, there are also dim and bright conditions, which make detection difficult. To this end, we set up contrast experiments on wheatear detection under different shooting conditions to compare the wheatear detection performance of these methods under a different light.

In the test set of this experiment, we selected 360 pictures with the same species and the number of wheatears in a single

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

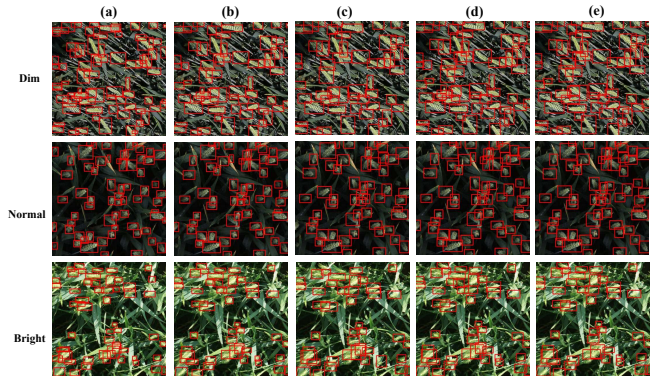


FIGURE 11. Wheatear detection results by different methods under different lighting conditions: (a) our method, (b) YoloV4, (c) method of Zhu et al., (d) method of Madec et al., (e) method of Yang et al..

TABLE 3. Wheatear detection results by different methods under different lighting conditions.

Light	Method	F1%			
		1	2	3	Average
normal	Our Method	96.13	96.15	96.17	**96.15**
	YoloV4 [7]	90.63	90.67	90.66	90.65
	Zhu et al. [10]	83.49	83.52	83.51	83.51
	Madec et al. [24]	91.34	91.36	91.32	91.34
	Yang et al. [26]	88.18	88.15	88.18	88.17
dim	Our Method	94.76	94.75	94.79	**94.77**
	YoloV4 [7]	90.13	90.15	90.14	90.14
	Zhu et al. [10]	80.73	80.75	80.71	80.73
	Madec et al. [24]	89.42	89.46	89.39	89.42
	Yang et al. [26]	84.34	84.37	84.35	84.35
bright	Our Method	94.61	94.58	94.59	**94.59**
	YoloV4 [7]	90.14	90.11	90.09	90.11
	Zhu et al. [10]	79.82	79.88	79.84	79.85
	Madec et al. [24]	88.18	88.14	88.17	88.16
	Yang et al. [26]	83.86	83.83	83.84	83.84
Average	Our Method	95.17	95.16	95.18	**95.17**
	YoloV4 [7]	90.30	90.31	90.30	90.30
	Zhu et al. [10]	81.35	81.38	81.35	81.36
	Madec et al. [24]	89.65	89.65	89.63	89.64
	Yang et al. [26]	85.46	85.45	85.46	85.46

picture ranging from 35 to 55, totally containing 17,382 wheatears. We divided the pictures into three groups according to the lighting conditions. There were 120 pictures in normal light, 120 in dim light, and 120 in bright light.

The experimental method is the same as the previous chapter, and the statistical results are shown in Table 3.

Table 3 shows that these methods can achieve good results under normal light conditions, but their detection performance is reduced when the light is dim or bright. Because the brightness is insufficient when the light is dim, and the texture of the wheatears is not clear when the light is bright, some wheatears will be directly eliminated, resulting in a decrease in f1-score. But on the whole, our method has more advantages and can adapt to different lighting scenarios. The detection results are shown in Figure 11.

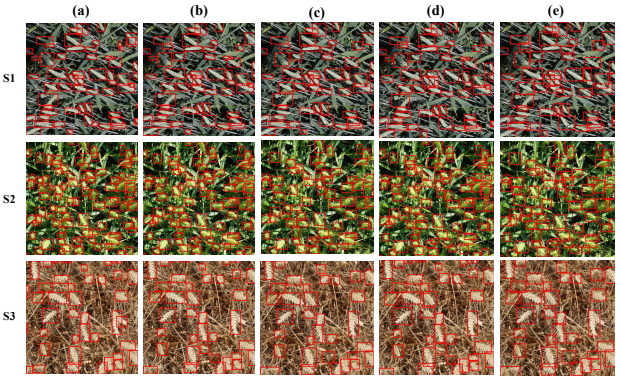


FIGURE 12. Detection results of different methods at various maturity stages of wheat: (a) our method, (b) YoloV4, (c) method of Zhu et al., (d) method of Madec et al., (e) method of Yang et al..

TABLE 4. Detection results of different methods at various maturity stages of wheat.

Wheat species	Method	F1%			
		1	2	3	Average
S1	Our Method	96.18	96.13	96.15	**96.15**
	YoloV4 [7]	90.15	90.08	90.13	90.12
	Zhu et al. [10]	82.84	82.86	82.83	82.84
	Madec et al. [24]	91.27	91.24	91.26	91.26
	Yang et al. [26]	83.84	83.86	83.86	83.85
S2	Our Method	94.67	94.64	93.68	**94.33**
	YoloV4 [7]	89.42	89.45	89.43	89.43
	Zhu et al. [10]	79.75	79.77	79.74	79.75
	Madec et al. [24]	88.91	88.95	88.93	88.93
	Yang et al. [26]	81.53	81.57	81.52	81.54
S3	Our Method	93.35	93.37	93.35	**93.36**
	YoloV4 [7]	88.95	88.97	88.93	88.95
	Zhu et al. [10]	78.24	78.21	78.25	78.23
	Madec et al. [24]	88.35	88.37	88.31	88.34
	Yang et al. [26]	81.27	81.24	81.23	81.25
Average	Our Method	94.73	94.71	94.39	**94.61**
	YoloV4 [7]	89.51	89.50	89.50	89.50
	Zhu et al. [10]	80.28	80.28	80.27	80.28
	Madec et al. [24]	89.51	89.52	89.50	89.51
	Yang et al. [26]	82.21	82.22	82.20	82.21

C. COMPARISON OF DETECTION RESULTS UNDER DIFFERENT MATURITY CONDITIONS

In the experiment in this chapter, the wheat maturity period is used as the control variable. Due to the different maturity periods, the color and appearance of wheatears vary as well. We established contrast experiments for wheatear detection at different maturity stages to compare the detection performance of various methods at various maturity stages.

In the test set in this experiment, we selected 360 pictures with a similar light environment and the number of wheatears in a single picture between 35 and 55, including 17,425 wheatears, which were divided into three groups according to the different maturity of wheat, each group contained 120 images.

The experimental method is the same as the previous chapter, and the statistical results are shown in Table 4.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3031896, IEEE Access

TABLE 5. Test results of different methods under the test set data.

Method	Avg-P	FPS
YoloV4 [7]	62.75	57.39
Zhu et al. [10]	53.64	8.53
Madec et al. [24]	71.32	12.46
Yang et al. [26]	56.46	21.62
Our Method	**77.68**	**72.13**

It can be seen from Table 4 that our method can maintain relatively stable performance in different maturity stages, and it still has a comparative advantage to adapt to the scenarios of different maturity stages. The detection results are shown in Figure 12.

In summary, the method of Zhu et al. [10] and method of Yang et al. [26] are difficult to detect in natural scenes such as dense wheatears and dim light, and the f1-scores scores are low; YoloV4 and Method of Madec et al. [24] have poor detection results when the wheatears are dense, and our method has advantages under various conditions. Three sets of comparative experiments show that our method can adapt to natural scenes and can detect wheatears more accurately.

The wheatear detection model based on the UAV platform not only requires accurate detection of wheatears but also requires the accuracy of the detection results and detection speed. To this end, we finally designed a set of contrast experiment and tested the models trained by different methods on the test set and calculated their Avg-P scores and FPS. The results are shown in Table 5.

It can be seen from Table 5 that although the method of Madec et al. [24] also has high accuracy, its speed is too slow. Our method has obvious advantages in speed and accuracy. Therefore, our method is more suitable as an embedded model of wheat detection UAV.

VI. CONCLUSION

Based on YoloV4, we have made improvements and proposed a robust wheatear detection method, which is suitable for UAV to detect wheatears in the field. This method can maintain excellent performance in natural scenes, including overlap, occlusion, light changes, different colors, and shapes. We have conducted a large number of comparative experiments, our method achieves 96.71%, 77.68%, and 72 on the three indicators of f1-score, avg-p, and FPS, respectively. The results show that our method is more suitable for wheatear detection based on UAV platform than existing methods. It has faster detection speed, higher detection accuracy, and better generalization ability. It can be used to estimate wheat density and spike size, and evaluate wheat health and maturity in large wheat fields by UAV. For our future research, we intend to explore a new architecture to further optimize wheatear detection (in terms of speed and accuracy of the bounding box). We also plan to extend the solution to the detection of crops with ears, such as the ear of rice. Due to the different shapes of crop ears, this brings new challenges.

REFERENCES

[1] V. Czymmek, R. Schramm and S. Hussmann, “Vision Based Crop Row Detection for Low Cost UAV Imagery in Organic Agriculture,” in *Proc. IEEE Int. Instrum. Meas. Tech. Conf. (I2MTC)*, Dubrovnik, Croatia, Mar. 2020, pp. 1-6.

[2] H. Xiang and L. Tian, "Development of a low-cost agricultural remote sensing system based on an autonomous unmanned aerial vehicle (UAV)", *Biosyst. Eng.*, vol. 108, no. 2, pp. 174-190, 2011.

[3] B. H. Y. Alsalam, K. Morton, D. Campbell and F. Gonzalez, "Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture", *Proc. IEEE Aerosp. Conf.*, pp. 1-12, Mar. 2017.

[4] H. Zheng, X. Zhou, T. Cheng, X. Yao, Y. Tian, W. Cao, and Y. Zhu, “Evaluation of a UAV-based hyperspectral frame camera for monitoring the leaf nitrogen concentration in rice.” In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, July. 2016, pp. 7350-7353.

[5] S. Yang, L. Hu, H. Wu, W. Fan and H. Ren, "Estimation Model of Winter Wheat Yield Based on Uav Hyperspectral Data," In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Japan, 2019, pp. 7212-7215.

[6] D. Stroppiana et al., "Rice yield estimation using multispectral data from UAV: A preliminary experiment in northern Italy," In *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Milan, 2015, pp. 4664-4667.

[7] A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” 2020, *arXiv:2004.10934*. [Online]. Available: https://arxiv.org/abs/2004.10934

[8] J. Redmon, A. Farhadi, “Yolov3: An incremental improvement,” 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[9] D. Zhang, Z. Wang, N. Jin, C. Gu, Y. Chen and Y. Huang, "Evaluation of Efficacy of Fungicides for Control of Wheat Fusarium Head Blight Based on Digital Imaging," in *IEEE Access*, vol. 8, pp. 109876-109890.

[10] Y. Zhu, Z. Cao, H. Lu, Y. Li, and Y. Xiao, “In-field automatic observation of wheat heading stage using computer vision,” *Biosyst. Eng.* vol. 143, pp. 28-41, Mar. 2016.

[11] J. A. Fernandez-Gallego, S. C. Kefauver, N. A. Gutiérrez, M. T. Nieto-Taladriz, and J. L. Araus, “Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images,” *Plant Methods* , vol. 2018, no. 1, pp. 22-34, Mar. 2018.

[12] J. Wu, G. Yang, X. Yang, B. Xu, L. Han and Y. Zhu, " Automatic counting of in situ rice seedlings from UAV images based on a deep fully convolutional neural network ", *Remote Sens.*, vol. 11, no. 6, pp. 691-710, Mar. 2019.

[13] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2001, vol. 1, pp. 511-518.

[14] P. Felzenszwalb, D. McAllester and D. Ramanan, "A Discriminatively Trained Multiscale Deformable Part Model", *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1-8, Jun. 2008.

[15] R. Girshick, "Fast R-CNN", *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1440-1448, 2015.

[16] J. Dai, Y. Li, K. He and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks", *Proc. Conf. Adv. Neural Inf. Process. Syst.*, pp. 379-387, 2016.

[17] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, “Focal Loss for Dense Object Detection,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999-3007.

[18] Z. Zhao, P. Zheng, S. Xu and X. Wu, “Object Detection With Deep Learning: A Review,” *IEEE Trans. Neural Netw. Learn. Syst*, vol. 30, no. 11, pp. 3212-3232, Nov. 2019.

[19] Q. Li, J. Cai, B. Berger, M. Okamoto, and S. J. Miklavcic, “Detecting spikes of wheat plants using neural networks with laws texture energy,” *Plant Methods*, vol. 13, no. 1, p. 83, Oct. 2017.

[20] C. Zhou, D. Liang, X. Yang, H. Yang, J. Yue, and G. J. Yang, “Wheat ears counting in field conditions based on multi-feature optimization and TWSVM,” *Frontiers Plant Sci.* vol. 9, p. 1024, Jul. 2018.

[21] Z. Akkus, A. Galimzianova, A. Hoogi, D. L. Rubin, and B. J. Erickson, "Deep learning for brain MRI segmentation: state of the art and future directions," *J. Digit. Imag.*,vol. 30, no. 4, pp. 449-459, 2017.

[22] J. A. Fernandez-Gallego, P. Lootens, I. Borra-Serrano, V. Derycke, G. Haesaert, I. Roldán-Ruiz, J. L. Araus, and S. C. J. T. P. J. Kefauver, "Automatic wheat ear counting using machine learning based on RGB UAV imagery," *The Plant Journal*, 2020.

[23] J. Ma, Y. Li, K. Du, F. Zheng, L. Zhang, Z. Gong, and W. Jiao, ''Segmenting ears of winter wheat at flowering stage using digital images and deep learning,'' *Comput. Electron. Agricult.*, vol. 168, Jan. 2020, Art. no. 105159.

[24] S. Madec et al., "Ear density estimation from high resolution RGB imagery using deep learning technique," *Agric. For. Meteorol*., vol. 264, pp. 225–234, 2019.

[25] D. Wang et al., "Combined Use of FCN and Harris Corner Detection for Counting Wheat Ears in Field Conditions," *IEEE Access*, vol. 7, pp. 178930-178941, 2019.

[26] Y. Yang, X. Huang, L. Cao, L. Chen and K. Huang, "Field Wheat Ears Count Based on YOLOv3," in *Proc. Int. Conf. Artif. Intell. Adv. Manuf. (AIAM)*, Oct. 2019, pp. 444-448.

[27] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779-788.

[28] C. Wang, H. Mark Liao, Y. Wu, P. Chen, J. Hsieh and I. Yeh, "CSPN et: A New Backbone that can Enhance Learning Capability of CNN," 2019, *arXiv:1911.11929.* [Online]. Available: http://arxiv.org/abs/1911.11929

[29] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, "Path Aggregation Network for Instance Segmentation," *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, jun. 2018, pp. 8759-8768.

[30] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, *Oct*. 2017, pp. 2961-2969.

[31] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 936-944.

[32] E. David, S. Madec, P. Sadeghi-Tehran, H. Aasen, B. Zheng, S. Liu, N. Kirchgessner, G. Ishikawa, K. Nagasawa, and M. Badhon, "Global Wheat Head Detection (GWHD) dataset: a large and diverse dataset of high resolution RGB labelled images to develop and benchmark wheat head detection methods," 2020, *arXiv:2005.02162.* [Online]. Available: http://arxiv.org/abs/2005.02162

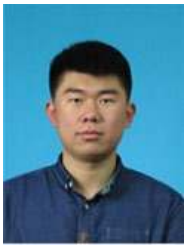
[33] T. DeVries, G. W. Taylor, "Improved regularization of convolutional neural networks with cutout" 2017, *arXiv:1708.04552.* [Online]. Available: http://arxiv.org/abs/1708.04552

[34] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo and J. Choe, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," 2019, *arXiv:1905.04899.* [Online]. Available: http://arxiv.org/abs/1905.04899.

[35] M. Tan, R. Pang, and Q. V. Le, ''EfficientDet: Scalable and efficient object detection,'' *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 10781–10790.

[36] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *in Proc. Workshop Challenges Represent. Learn. (ICML)*, vol. 3, 2013, pp. 2-7.

MING-XIANG HE is currently a Professor with the College of Computer Science and Engineering, Shandong University of Science and Technology, China. He is currently a member of the National Virtual Simulation Experiment Center of Shandong University of Science and Technology. His current research interests include image processing, artificial intelligence, database system



PENG HAO is currently pursuing the master's degree in software engineering from Shandong University of Science and Technology, China. His current research interests include deep learning and object detection and text recognition.

YOU-ZHI XIN is currently pursuing the master's degree in software engineering from Shandong University of Science and Technology, China. His current research interests include deep learning and object detection and text recognition.