



OPEN A generalized model for accurate wheat spike detection and counting in complex scenarios

Changji Wen¹✉, Zhenyu Ma¹, Junfeng Ren¹, Tian Zhang¹, Long Zhang¹, Hongrui Chen¹, Hengqiang Su¹, Ce Yang², Hongbing Chen^{1,4} & Wei Guo³✉

Wheat is a crucial crop worldwide, and accurate detection and counting of wheat spikes are vital for yield estimation and breeding. However, these tasks are daunting in complex field environments. To tackle this, we introduce RIA-SpikeNet, a model designed to detect and count wheat spikes in such conditions. First, we introduce an Implicit Decoupling Detection Head to incorporate more implicit knowledge, enabling the model to better distinguish visually similar wheat spikes. Second, Asymmetric Loss is employed as the confidence loss function, enhancing the learning weights of positive and hard samples, thus improving performance in complex scenes. Lastly, the backbone network is modified through reparameterization and the use of larger convolutional kernels, expanding the effective receptive field and improving shape information extraction. These enhancements significantly improve the model's ability to detect and count wheat spikes accurately. RIA-SpikeNet outperforms the state-of-the-art YOLOv8 detection model, achieving a competitive 81.54% mAP and 90.29% R². The model demonstrates superior performance in challenging scenarios, providing an effective tool for wheat spike yield estimation in field environments and valuable support for wheat production and breeding efforts.

Keywords Wheat spikes, Detection and counting, RIA-SpikeNet, Crop phenotype, Complex scenarios

More than 30% of the world's population lives on wheat, which plays an important role in human diet¹⁻³. With the growth of the population and the improvement of living standards, the demand for wheat continues to increase. In addition, wheat production has been affected by climate change and the reduction of cultivated land⁴⁻⁶. Therefore, it is very important to increase wheat yield through breeding. The large-scale, multi-category, and multi-morphological phenotypic information provided by wheat spike numbers is of great significance for optimizing variety selection and improving yield^{7,8}. This data allows for a more comprehensive understanding of the wheat growth process and can provide valuable insights into the genetic and environmental factors that affect yield. By analyzing this data, researchers can identify specific traits that are associated with high-yielding varieties and develop targeted breeding programs to enhance these traits. Therefore, how to effectively count wheat spikes is particularly critical. Traditional manual statistical methods are time-consuming and inefficient. Developing an efficient and automatic method for wheat spike detection and counting per unit area is necessary and beneficial. Of particular significance is the multi-category, multi-class wheat spike detection. This is primarily aimed at achieving accurate classification and identification of wheat varieties, thereby helping agricultural scientists and farmers to better understand the characteristics and performance of different wheat varieties for proper cultivation and management. This helps to improve the yield and quality of wheat and promotes sustainable agriculture⁹.

Due to the popularity of camera platforms, image-based wheat spike detection is a potential solution to replace tedious manual observation. To develop efficient and automatic wheat spike detection algorithms, a large and diverse wheat spike dataset is necessary. Several state-of-the-art wheat spike detection datasets are summarized here. Hasan et al.¹⁰ constructed the SPIKE dataset using ground-based equipment and cameras. The acquired images capture wheat spikes in a plot area of 0.8 m × 0.8 m. The dataset consists of 335 images, including 10 wheat varieties at different growth stages, which contain about 25,000 wheat spikes. Zhao et al.¹¹ took a wheat spike dataset through Unmanned Aerial Vehicle (WSUAV). They divided the three wheat fields

¹College of Information and Technology, Jilin Agricultural University, Changchun, China. ²College of Food, Agricultural and Natural Resource Sciences, University of Minnesota, Twin Cities, Minnesota, MN, USA.

³Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo, Japan. ⁴Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen, China. ✉email: changjiw@jlau.edu.cn; guowei@g.ecc.u-tokyo.ac.jp

into acquisition scenarios of different periods, densities and resolutions. Then, in nine different scenes, 3171 images, including 82,873 wheat spikes, were obtained. To address the limitations of the wheat spike dataset, David and Guo et al. proposed two Global Wheat Head Detection (GWHD) Dataset in 2020 and 2021^{12,13}. They use multiple devices to capture images. Then, wheat spikes are marked using a web platform. The GWHD2021, which is an amplification and optimization of GWHD, contains 47 types of wheat spikes from 12 countries, with a total of 275,187 wheat spikes from 6515 images. The dataset has the advantages of large-scale, multi-category, and multi-morphological phenotypic data.

Based on the comparison of several datasets in Table 1, it is evident that the GHWD2021 dataset stands out with its larger number of images and diverse categories. In contrast, the SPIKE dataset has a limited number of images and categories, which makes it challenging to meet the requirements of deep learning that demand a large volume of data. While the WSUAV dataset and GWHD2020 dataset have a larger data volume, they lack diversity in their data. A robust deep learning model for all situations requires a dataset of images covering a wide range of genotypes, seeding densities and patterns, plant states and stages, and collection conditions. Similarly, the plant phenotyping community recognizes the importance of having access to plant- or crop-specific datasets. The phenotyping community also recognizes the importance of having access to plant- or crop-specific datasets. Considering its characteristics, the GWHD2021 dataset appears to be more suitable for wheat spike detection.

In recent decades, with the development of computer technology, image processing and machine learning techniques have been increasingly applied to extract crop phenotype parameters^{14,15}. These methods have made some contributions to wheat yield estimation. However, there are still limitations such as requiring too much manual intervention and weak model generalization ability. Deep learning models developed in recent years have shown advantages over traditional image processing and machine learning methods in most visual tasks¹⁶. Deep learning has the advantage of automatically learning features from big data, instead of manually designed features, and has high accuracy¹⁷. Among them, deep learning-based object detection algorithms have been greatly improved, and various two-stage and one-stage algorithms have emerged. With the development of deep learning algorithms, their application in the field of agriculture is becoming more widespread. Wang et al.¹⁸ improved on the EfficientDet-D0 target detection network. They added a convolutional block attention module¹⁹ (CBAM) to the network to refine the features and then enhanced the images using data enhancement simulating wheat occlusion. Thus, the wheat occlusion problem is solved and the final counting accuracy reaches 94% on the GWHD2020 dataset. Sun et al.²⁰ improved the Faster R-CNN network. The network uses Augmented Feature Pyramid²¹ (AugFPN) for adaptive aggregation of raw information, which solves the problem of insufficient wheat spikes. Then they adjusted the union (IoU) threshold to effectively remove the interference in complex backgrounds and improve the wheat spike localization accuracy. On the GWHD2020 dataset, the network averages an error rate of only 3.7%, and Ap reaches 95.17%. Khan et al.²² proposed a deep learning framework combining DenseNet and InceptionNet to capture multi-scale features, along with an attention module to address inter-class similarity and intra-class variability in plant disease detection. This framework enhances the model's ability to focus on discriminative features while ignoring irrelevant information. Wen et al. proposed SpikeRetinaNet for the detection and counting of wheat spikes. The network solves the problem of shading in the field by using a weighted bidirectional feature pyramid network (BiFPN) and soft non-maximal suppression (Soft-NMS). SpikeRetinaNet achieved an mAP of 92.62% on the GWHD2020 dataset²³. There has been some study using GWHD2021 as a dataset. The study by Han and zang et al.^{2,24,25} used GWHD2021 for wheat spike detection and got good results. Liu et al.¹ proposed Dynamic Color Transformation (DCT). The DCT model changes the color channel of the input image, which can significantly reduce false positives and improve the detection performance. It was used on the YOLOV4²⁶ network and obtained a average domain accuracy (ADA) of 69.5% on the GWHD2021 dataset, which is the runner-up of GWC2021. Qing et al.²⁷ and Zhou et al.⁴¹ respectively proposed the YOLO-FastestV2 model and MW-Swin Transformer for the detection and counting of wheat spikes. As a result of these optimizations, the model's feature extraction capabilities and overall accuracy were significantly improved achieving relatively outstanding detection results on the incremental wheat dataset based on GWHD2021. Meng et al.²⁸ proposed the YOLOv7-MA model to address the challenges of overlapping wheat spikes and small target sizes in complex backgrounds. By incorporating a micro-scale detection layer and convolutional block attention module, the model enhances spike detection and reduces background interference, achieving a mean average precision (MAP) of 93.86% and a detection speed of 35.93 FPS on the Global Wheat Head Dataset 2021. A deep learning framework was proposed by Khan et al.²⁹, combining DenseNet and InceptionNet to capture multi-scale features, along with an attention module designed to focus on the most salient characteristics. Experimental results on public datasets demonstrated the framework's effectiveness in achieving high accuracy in plant disease detection, overcoming challenges related to visual similarity and variability. Despite advancements, current deep learning models for agricultural tasks face several challenges. Many require significant manual intervention and have weak generalization, limiting their real-world applicability. Their complex architectures and high computational demands hinder practical use

Data set	Number of pictures	Number of wheat spikes	Number of classes	Average number of wheat spikes in the picture
SPIKE	335	25,000	10	74.62
WSUAV	3171	82,873	9	26.13
GWHD2020	4700	193,634	11	41.19
GWHD2021	6515	275,187	47	42.23

Table 1. The numerical values of each item in the dataset.

in resource-limited environments. Adaptability to new conditions and varying datasets remains an issue, with inconsistent accuracy across different contexts. Furthermore, models struggle to handle complex backgrounds effectively, such as occlusion and shading, requiring further refinement.

While the aforementioned works have made progress in wheat spike detection, they lack analysis on various challenging scenarios. Correspondingly, the datasets also face diverse visual challenges. As shown in Fig. 1, it illustrates several challenges, which are specified as follows:

1. Effect of light variation: Due to weather fluctuations and variations in shooting angles, the lighting conditions in the GHWD2021 dataset differ significantly. This leads to unclear and challenging detection of wheat spikes, particularly under dim lighting.
2. Similar color gamut between wheat spikes and background: In certain growth stages, the color gamut of wheat spikes closely resembles that of the background. This similarity poses difficulties in accurately detecting wheat spikes, especially when they are green during the post-flowering stage or entirely yellow during the ripening stage.
3. Shading and overlapping problems: Natural conditions can cause wheat spikes to overlap or be shaded by leaves. When severe, these issues make it challenging for the model to correctly identify and detect wheat spikes.
4. Complex background: The presence of a complex background adds diversity to the dataset. However, it also introduces challenges in distinguishing wheat spikes based on texture features alone, as the complexity can cause confusion.

To address the aforementioned environmental challenges, this paper presents RIA-SpikeNet, an efficient model for accurate wheat spike detection and counting, by improving YOLOX. Our model effectively tackles various challenges encountered in detecting wheat spikes in complex field environments. These challenges include the similarity between wheat spikes and the background color gamut, imbalances in positive and negative samples, difficulties in sample learning, and issues related to occlusion and overlapping of leaves and spikes. The main contributions of this study are as follows:

1. We propose a head detection approach using an implicit decoupling method. By leveraging the learning of implicit semantic knowledge to fill into the implicit model, our model acquires valuable contextual information, enabling more accurate spike detection.
2. We employ the asymmetric loss and emphasize the learning of samples by adjusting the gradients of positive and negative sample losses. By enhancing the weights of positive samples and difficult samples' learning weights, we strengthen the learning ability of our model, thus resolving the problem of imbalanced positive and negative samples and sample learning difficulties under complex conditions.
3. By employing RepLKNet as the backbone, this model leverages a larger convolutional kernel and reparameterization structure, thereby possessing a larger effective receptive field and shape information. The larger effective receptive field enables the capture of more global feature information, preventing the loss of valuable information and enhancing the extraction of information for both the spike and background. Additionally,

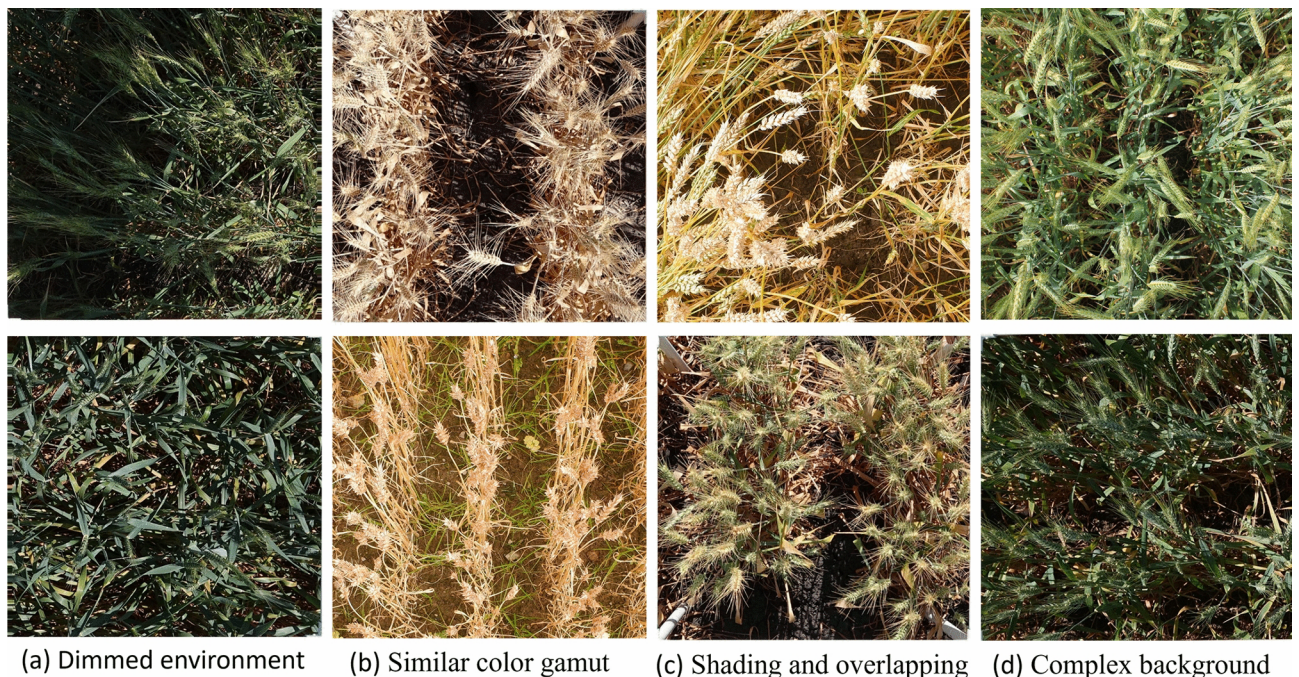


Figure 1. Wheat spikes under different difficult conditions.

the increased shape information facilitates better identification of spikes from backgrounds with complex texture features.

- We conduct extensive experiments to demonstrate the superiority of our RIA-SpikeNet network. The proposed RIA-SpikeNet achieves a mAP of 81.54% and an R2 of 90.29%, surpassing state-of-the-art detection models such as YOLO8 by 2.46% and 2.76% respectively. These results demonstrate the outstanding performance of our model in detecting and counting wheat spikes with diverse categories and variations.

Materials and methods

Data acquisition set processing

The wheat spike dataset, GWHD2021, is from 12 countries. It covers genotypes from different regions of the world, such as Europe, Africa, Asia, Australia and North America. Those from Asia include Nanjing, China and Tokyo, Japan. The wheat spike dataset consists of RGB images that were captured using different ground-based phenotypic platforms and cameras, such as Spidercam, Gantry, Cart, handheld, and Tractor, including four growth stages of wheat, post-flowering, filling stage, filling-maturity and maturity. The images were captured at a height ranging from 1.8 m to 3 m, and the focal length of the acquisition camera varied from 10 to 50 mm. The dataset has been classified based on specific criteria, where a group of images acquired with the same equipment in the same area is defined as a class. The images that were captured undergo a manual screening process to remove any blurred or invalid images. After this, they are segmented into images that have 1024×1024 pixels. Each of these processed images is labeled for each wheat spike using a web-based labeling platform. As shown in Fig. 2, the dataset is divided into different stages and the labeling effect.

GWHD2021 contains 6,515 images of wheat spikes, with each image containing 20 to 60 wheat spikes. There are 47 classes of wheat spikes in the dataset and a total of 275,187 wheat spike labels. The wheat spike dataset underwent data cleaning and division before conducting further research. In the data cleaning process, 128 unlabeled images were removed from the dataset. Next, we removed any classes with less than 10 images, resulting in the final wheat spike dataset consisted of 6,374 images that were divided into 45 classes. The specific information of the dataset is shown in Fig. 3, which includes the country, sample size, and collection tools. These data were divided into training and test sets according to 8:2.

In order to enhance the generalization of the model, the images were augmented with Mosaic before training. Mosaic data augmentation is a technique used during model training to improve its generalizability by combining multiple images into a single training batch³⁰. The data augmentation process involves the following steps: First, a mosaic canvas is generated. Then, four images are randomly selected and stitched together, corresponding to the top-left, top-right, bottom-left, and bottom-right quadrants of the mosaic canvas, respectively. Next, the images undergo random transformations, including flipping, resizing, and color jittering. Finally, the images are subjected to coordinate transformations, including cropping and bounding box adjustments, to ensure that they do not exceed the boundaries of the mosaic canvas. This technique involves combining four training images into

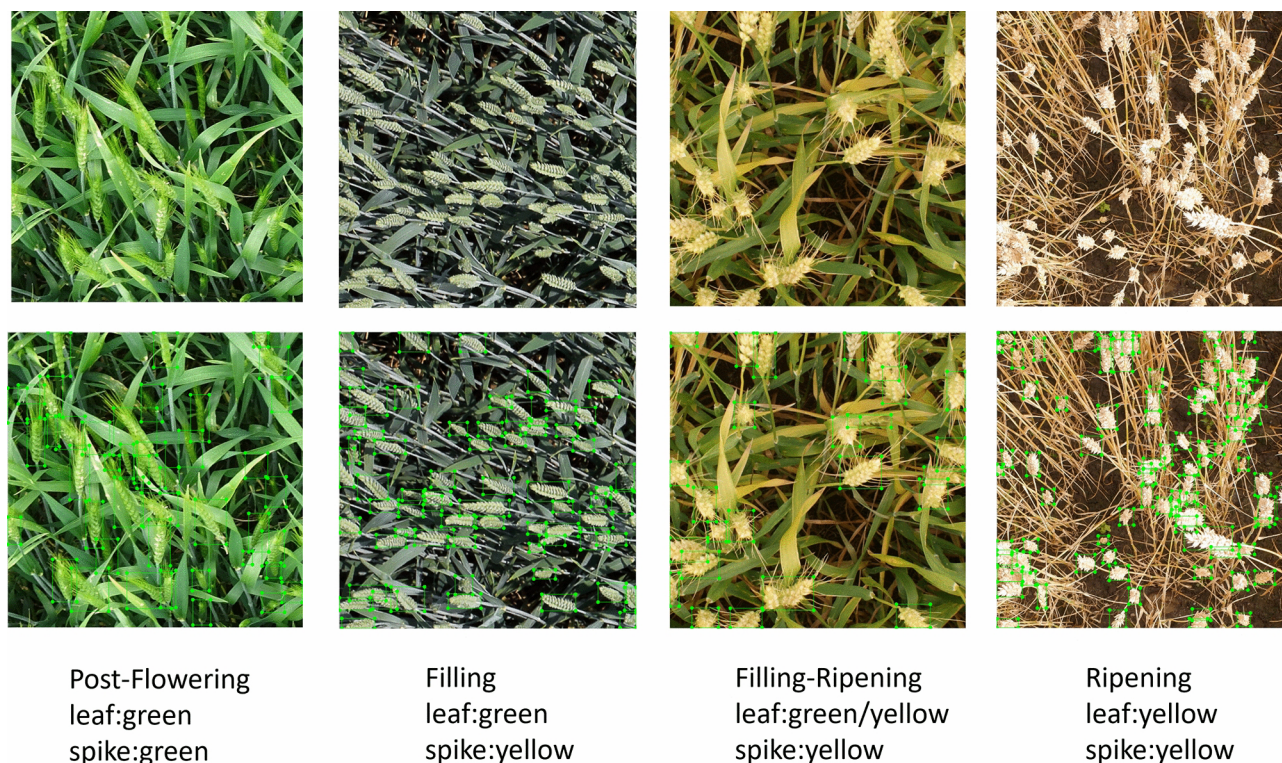


Figure 2. Wheat spike annotation at different stages.

Class	Number of images	Country	Development stage	Platform
ETHZ_1	747	Switzerland	Filling	Spidercam
Rres_1	432	UK	Filling - Ripening	Gantry
ULiège-GxABT_1	30	Belgium	Ripening	Cart
NMBU_1	82	Norway	Filling	Cart
NMBU_2	96	Norway	Ripening	Cart
Arvalis_1	66	France	Post-flowering	handheld
Arvalis_2	401	France	Filling	handheld
Arvalis_3	588	France	Filling - Ripening	handheld
Arvalis_4	204	France	Filling	handheld
Arvalis_5	402	France	Filling	handheld
Arvalis_6	160	France	Filling - Ripening	handheld
Arvalis_7	24	France	Filling - Ripening	handheld
Arvalis_8	20	France	Filling - Ripening	handheld
Arvalis_9	32	France	Ripening	handheld
Arvalis_10	58	France	Filling	handheld
Arvalis_11	60	France	Filling	handheld
Arvalis_12	29	France	Filling	handheld
Inrae_1	176	Canada	Filling - Ripening	handheld
Usask_1	200	US	Filling - Ripening	Tractor
KSU_1	100	US	Post-flowering	Tractor
KSU_2	100	US	Post-flowering	Tractor
KSU_3	95	US	Filling	Tractor
KSU_4	60	US	Ripening	Tractor
Terraref_1	133	US	Ripening	Gantry
Terraref_2	86	US	Filling	Gantry
CIMMYT_1	66	Mexico	Post-flowering	Cart
CIMMYT_2	74	Mexico	Post-flowering	Cart
CIMMYT_3	52	Mexico	Post-flowering	Cart
Utokyo_1	538	Japan	Ripening	Cart
Utokyo_2	456	Japan	Ripening	Cart
Utokyo_3	114	Japan	multiple	handheld
Ukyoto_1	58	Japan	Post-Flowering	handheld
NAU_1	20	China	Post-Flowering	handheld
NAU_2	100	China	Post-Flowering	Cart
NAU_3	100	China	Filling	Cart
UQ_1	21	Australia	Post-flowering	Tractor
UQ_4	29	Australia	Filling	Tractor
UQ_5	29	Australia	Filling - Ripening	Tractor
UQ_6	28	Australia	Filling - Ripening	Tractor
UQ_7	17	Australia	Ripening	handheld
UQ_8	41	Australia	Ripening	handheld
UQ_9	33	Australia	Filling - Ripening	handheld
UQ_10	106	Australia	Filling - Ripening	handheld
UQ_11	84	Australia	Post-flowering	handheld
ARC_1	30	Sudan	Filling	handheld

Fig. 3. Detailed information of the dataset.

a single mosaic image. This technique helps to increase the diversity and complexity of the training data, leading to improved model performance and better generalization capabilities.

Methods

This research focuses on designing a high-performance model for the wheat spike detection and counting. To achieve this goal, we have selected YOLOX as the baseline architecture for our model. YOLOX is a state-of-the-art object detection model that offers high accuracy and fast processing speed. It is based on YOLOv3³¹

and uses the anchor-free³² approach, decoupled headers, and SimOTA label assignment strategy. Additionally, YOLOX employs YOLOv5 backbone and network extension rules to construct networks of different sizes. Since the position, size, and orientation of wheat spikes can vary, an anchor-free design is essential for detecting them accurately. YOLOX's anchor-free approach makes it an ideal choice for our research. Therefore, we chose YOLOX as our baseline.

Several strategies have been adopted to improve model performance considering the complex scenarios of field wheat spike detection and counting. We improved the detection head, confidence loss and backbone network respectively, and finally obtained RIA-SpikeNet. The overall model structure of RIA-SpikeNet is shown in Fig. 4.

As depicted in Fig. 3, the Mosaic technique first stitches together the four input images into a single image. This combined image is then processed by the RepLKNet, which serves as the backbone network. RepLKNet is chosen for its ability to extract shape information and larger effective fields, making it well-suited for object detection tasks like wheat spike detection. The backbone network begins by capturing shallow image features through four convolutional operations. It then passes through four Stage blocks, consisting of RepLK and ConvFFN blocks, for deeper feature extraction. The Transition block downsamples the features to create features at different scales. After feature extraction from Stage2 to Stage4, these blocks are output into PANet to fuse the shallow and deep features. Initially, up-sampling fuses the deep features with the shallow features, resulting in C3, C4, and C5, which are further downsampled and fused again with the deep features to produce P3, P4, and P5. Finally, the output features are forwarded to the implicit decoupling detection head (imHead) to generate network prediction results. The detection head decouples the features at different scales and adds an implicit model for refining the feature. The ultimate features are separated into classification branch (cls), regression frame branch (reg) and confidence branch (obj) outputs. Implicitly decoupling the detection head allows for a better understanding of the context of the objects in the image, enabling the model to distinguish between similar objects and reducing the likelihood of misclassifying objects or detecting false positives. During training, the predictions are assigned labels by SimOTA, and losses are calculated using the cross-entropy loss function, the asymmetric loss function, and the IOU loss function, respectively. The Asymmetric Loss is adopted as the confidence loss, which assigns different penalties for false positives and false negatives helping the model to focus more on correctly identifying difficult or rare samples. This is particularly important in scenarios where the distribution of positive and negative samples is imbalanced, such as in the case of field wheat spike detection and counting. During prediction, the results are non-maximum suppressed (NMS)³⁵. To remove redundant predictions, and the predicted wheat spikes are displayed with bounding boxes. We show the algorithmic flow of the model in Fig. 5.

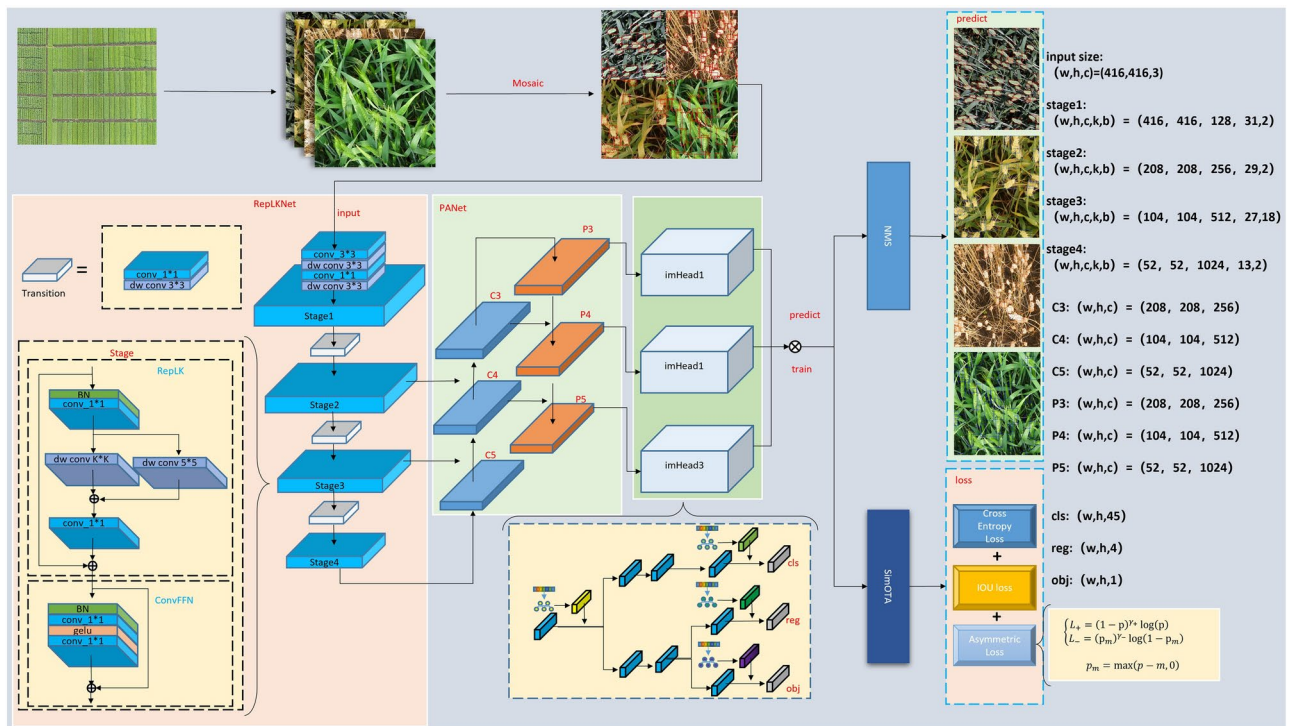


Fig. 4. RIA-SpikeNet model structure framework diagram. Mosaic is the data enhancement method. RepLKNet is used as the backbone network. PANet for feature fusion. The imHead is the implicit decoupling detection head. w , h and c denote the width, height and number of channels of the feature map. C3, C4, and C5 represent the input feature maps, while P3, P4, and P5 represent the output feature maps. The conv denotes normal convolution, dw_conv denotes depth separable convolution, gelu³³ is the activation function, and BN is batch normalization³⁴.

Algorithm 1:

Input: Training set, test set, hyperparameters
Output: loss, detection box, category, number of detections

- 1: **If training then :**
- 2: Mosaic the train set for data enhancement
- 3: Inputting the train set into the model
- 4 **else :**
- 5: Inputting the test set into the model
- end if
- 6: **For Epoch—0 to 300 do :**
- 7: Extract features using RepLKNet backbone network
- 8: The last three feature layers are output:
 - C3: $(w, h, c) = (208, 208, 256)$
 - C4: $(w, h, c) = (104, 104, 512)$
 - C5: $(w, h, c) = (52, 52, 1024)$
- 9: Different features of the fusion output using FPANet:
 - P3: $(w, h, c) = (208, 208, 256)$
 - P4: $(w, h, c) = (104, 104, 512)$
 - P5: $(w, h, c) = (52, 52, 1024)$
- 10: The output features go to the decoupling detection head separately, and for decoupling the outputs are classification, regression frames and confidence results:
 - cls: $(w, h, 45)$
 - reg: $(w, h, 4)$
 - obj: $(w, h, 1)$
- If training :**
- 11: The results are selected by simOTA for the candidate box, and the prediction results corresponding to the target box are selected.
- 12: Calculate the loss according to the formula:
 - Cross Entropy Loss
 - IOU loss
 - Asymmetric Loss
- 13: Calculate the gradient and propagate backwards
- 14: **Return loss**
- 15: **else :**
- 16: The test results were obtained after the NMS race selection
- 17: **Return detection box, category, number of detections**
- end if
- end for

Fig. 5. RIA-SpikeNet Model Algorithm Flow. w , h and c denote the width, height and number of channels of the feature map. C3, C4, and C5 represent the input feature maps, while P3, P4, and P5 represent the output feature maps.

Implicit decoupling detection head

The color gamut similarity in the natural environment can have a significant impact on the accuracy of wheat ear recognition. To address this challenge, we have redesigned the detection head and proposed an implicitly decoupled detection head. We have added implicit models to the original decoupled detection head to incorporate implicit knowledge³⁶, which refers to semantic information that cannot be observed through pictures. By adding

implicit models to the detection head, we can incorporate more semantic information to refine the features and improve the accuracy of wheat ear detection. Implicit semantic knowledge can be considered as auxiliary information related to the shape, color, texture, and other fundamental aspects of the target object. It includes features such as the shape characteristics of wheat spikes and the surrounding morphology of other plants. The modeling process of the implicit model is as follows.

$$\mathcal{Y} = \mathcal{F}_\theta(\xi) + \varepsilon \quad (1)$$

$$\mathcal{Y} = \mathcal{F}_\theta(\xi) + \mathcal{Z}(\ddagger) + \varepsilon \quad \text{Minimize } \mathcal{Z}(\ddagger) + \varepsilon \quad (2)$$

The traditional network training objective function is represented by Eq. 1, where ξ is the input features, \mathcal{F}_θ is the function expression, and ε is the bias term. To improve the accuracy of the predictions by incorporating additional semantic information from the implicit knowledge, we model the implicit knowledge using vectors and add it to the traditional function using an addition operation. This can be expressed as shown in Eq. 2, where \mathcal{Z} denotes vector construction and \ddagger is the value of the vector representing the implicit knowledge. The value of \ddagger is initialized randomly and is obtained through the learning of forward and backward propagation. By adding the implicit model to the detection head, we can refine the features and improve the accuracy of the predictions for detecting wheat ears in complex field scenes.

The detection head is shown in Fig. 6. We introduced an implicit model at the input and output of the detection head to obtain an implicitly decoupled detection head. Specifically, the implicit decoupling detection head first applies the convolution operation to the input features and then adds the implicit model (\mathcal{Z}). Subsequently, the features are decoupled into three branches by multiple convolutions, namely classification branch (cls), regression boxes branch (reg), and confidence branch (obj). Finally, the implicit model (\mathcal{Z}) is added again to the output of each branch. The dimension of the implicit model added at different positions of the detection head varies and is related to the input feature map. The initial construction of implicit semantic knowledge takes the form of random vectors, which are then refined during the training process to capture implicit semantic knowledge related to the shape characteristics of wheat spikes and other surrounding morphological information. However, before adding the implicit model, it needs to be expanded to the input feature size. The physical meaning of adding an implicit model is to incorporate the learned implicit weights into the features and refine them. By adding the implicit model, the information of features can be enriched and refined. Therefore, the implicit decoupling detection head can better distinguish objects of similar color gamut and achieve the purpose of improving detection accuracy.

Asymmetric loss function

In wheat detection, there are both correct and incorrect predictions. Positive samples refer to correct predictions, while negative samples refer to incorrect predictions. Detecting wheat spikes in complex environments generates

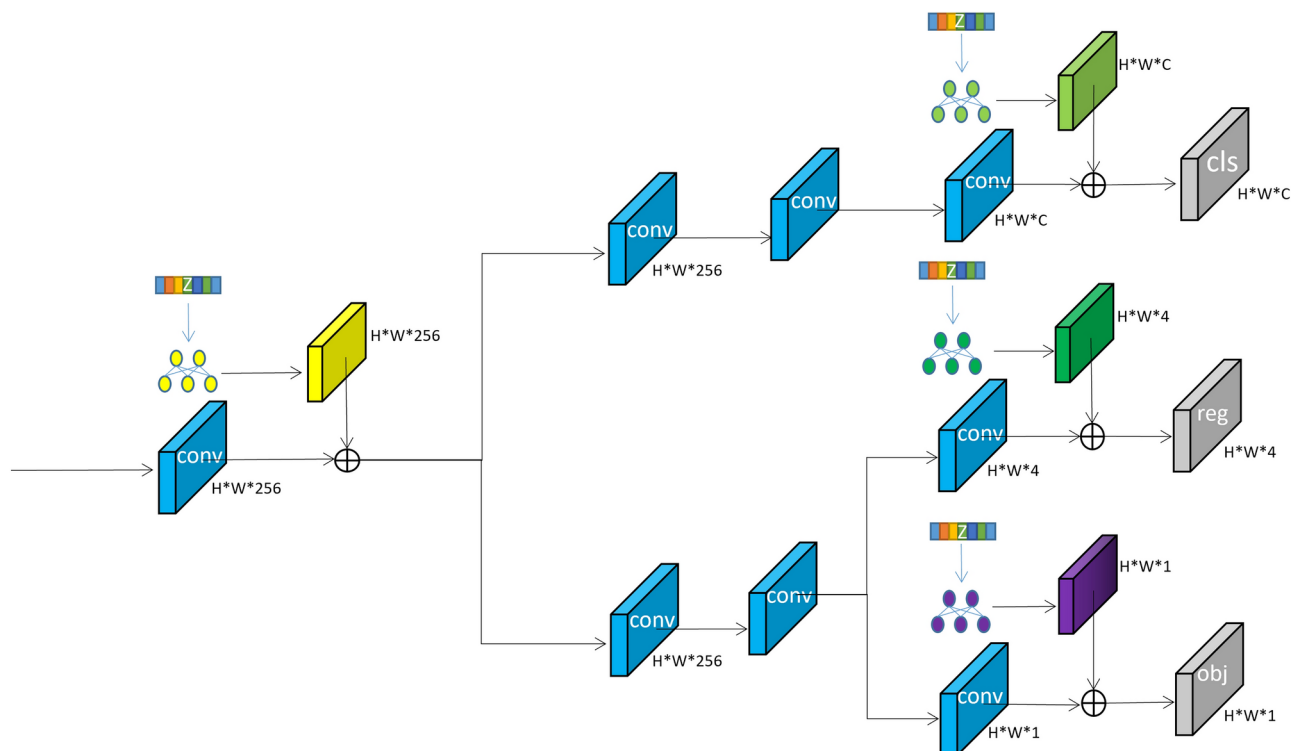


Fig. 6. Implicit decoupling detection head.

more negative and difficult samples during the detection process. When the number of negative samples is too large, it can inhibit the learning of positive samples and affect the detection accuracy. To address this issue, we suggest using Asymmetric Loss instead of cross-entropy loss as the confidence loss. Equation (3) is the cross-entropy loss. In the formula p is the prediction probability, L_+ indicates the positive sample loss, and L_- is the negative sample loss. This loss is calculated using the logarithmic function for both positive and negative sample losses. Asymmetric loss uses different weights to adjust the contributions of positive and negative samples, as shown in Eq. (4). In the formula γ_- and γ_+ are the focusing parameters, and m is the hard threshold.

$$CE = \begin{cases} L_+ = \log(p) \\ L_- = \log(1 - p) \end{cases} \quad (3)$$

$$ASL = \begin{cases} L_+ = (1 - p)^{\gamma_+} \log(p) \\ L_- = (p_m)^{\gamma_-} \log(1 - p_m) \end{cases} \quad p_m = \max(p - m, 0) \quad (4)$$

We mainly regulated by two ways of loss. The first way is to use the decoupled focus levels for positive and negative samples, and set γ_- and γ_+ as the parameters of positive and negative focus, respectively, as in Eq. (5). The size of the weights is adjusted by controlling γ_- and γ_+ , and the lower the parameter's value, the higher the contribution of the loss. In this experiment, the values of γ_- and γ_+ are taken in the range of [0, 1, 2, 3, 4, 5]. Since we want to enhance the contribution of positive samples, the focusing parameter is set to $\gamma_- > \gamma_+$. This helps to learn more meaningful features from the positive samples.

$$\begin{cases} L_+ = (1 - p)^{\gamma_+} \log(p) \\ L_- = p^{\gamma_-} \log(1 - p) \end{cases} \quad (5)$$

The second way is to add a hard threshold m to the loss of negative samples, which is used to reduce the gradient generated by the easily detectable negative samples in the negative samples. As shown in Eq. (6), we assign the loss of negative samples with prediction probability less than m to 0. By doing so, we eliminate the effect of simple negative samples, thus emphasizing more on the gradient of positive samples. The m is set not only to reduce the gradient of negative samples, but also to enhance the gradient of difficult samples.

$$L_- = (p_m)^{\gamma_-} \log(1 - p_m) \quad p_m = \max(p - m, 0) \quad (6)$$

RepLKNet backbone network

In complex scenes with shadows and overlapping foliage, texture features are particularly complex and wheat spikes are difficult to detect. To overcome the problems of complex backgrounds, overlapping and occlusion, we introduced the RepLKNet as the backbone network, which has a larger effective receptive field and shape information³⁷. The effective receptive field is a metric proposed by Luo et al.³⁸ to describe the effect of each input pixel in the receptive field on the output of the n -layer cells in the network. The network structure is illustrated in Fig. 7 and consists of three blocks: Stem block, Stage block and Transition block. By using RepLKNet, which has a larger effective receptive field, the model can better extract features from the input images, especially in complex scenes with shadows and overlap with leaves, where small wheat spikes are difficult to be detected. In addition, the complex background in the wheat growing environment makes the texture features particularly complex. The network has more shape information, which makes it possible to better detect wheat spikes by shape features in complex textured environments. These improvements help overcome the problems of complex backgrounds, overlap and occlusion, and improve the accuracy of the model in these scenes.

Results

Model equipment and parameters

The model was implemented using the Python and the Pytorch framework. The experiments were conducted on a machine with an Intel(R) Xeon(R) CPU, NVIDIA GPU GeForce RTX 2080Ti, and 16 GB of RAM. The operating system used was Ubuntu 18.04, and the experimental environment was Pytorch-1.7.0, cuda-10.1, cudnn-7.6.5.

We split the dataset into a training set of 5082 images and a test set of 1292 images. The training set was then augmented using Mosaic and fed into the model for training. The test set, on the other hand, was kept as is for final testing of the model. The model was trained using the Adam optimizer. The learning rate decay strategy was set to a step decay, decaying every 100 epochs with a decay rate of 0.1. The initial learning rate was set to 0.0001. The input image size to the model was $416 \times 416 \times 3$. The model was trained for a total of 300 epochs, with a batch size of 2.

Evaluation indicators

In this study we used different evaluation metrics to assess the detection and counting of wheat spikes. We use precision (P), recall (R), F1 score, and mean accuracy (mAP) as the main evaluation metrics for wheat spike detection. The prediction results are classified into four categories: true negative (TP), false positive (FP), true negative (TN) and false positive (FN)³⁹. Precision and Recall are calculated based on the above four categories, and the formula is shown in Eqs. 7, 8. Precision is the ratio of the number of correct predictions to the number of all predictions. Recall is the ratio of the number of correct predictions to the number of all labels⁴⁰. F1 score is a useful measure to observe the robustness of the model, calculated using Precision and Recall, and is the summed average of Precision and Recall, as shown in Eq. 9. AP is the the area under the Precision and Recall curves and

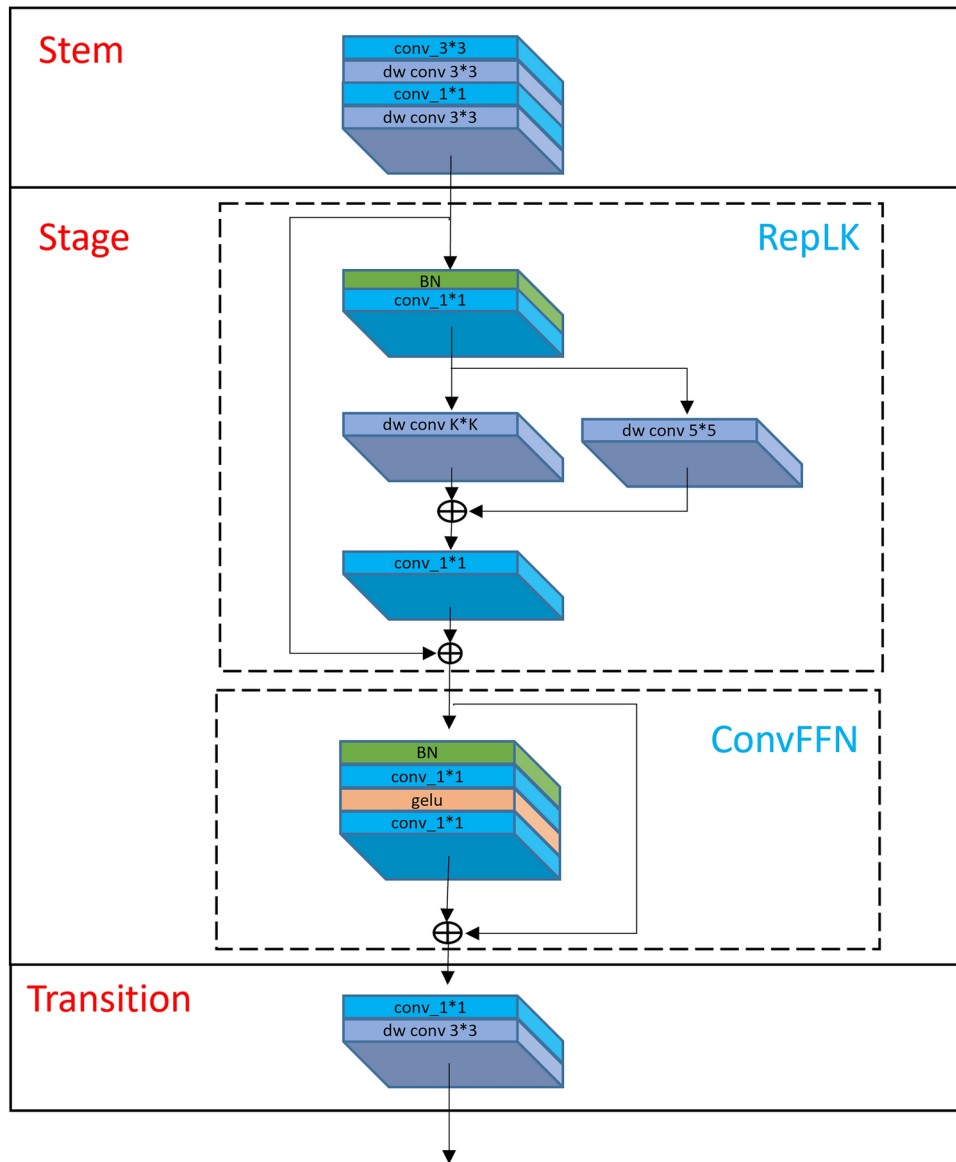


Fig. 7. RepLKNNet Network Architecture Diagram.

is a measure of the overall performance of the object detection, as shown in Eq. 10. The mAP is the average of AP over all classes, as shown in Eq. 11.

$$precision(P) = \frac{TP}{TP + FP} \tag{7}$$

$$recall(R) = \frac{TP}{TP + FN} \tag{8}$$

$$F1 \text{ score} = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

$$AP_K = \int_0^1 P(R_K) dR_K \tag{10}$$

$$mAP_K = \frac{1}{45} \sum_{k=1}^{45} AP_K \tag{11}$$

For count evaluation, we chose to use the mean average error (MAE), root mean square error (RMSE), and correlation coefficient (R^2) as evaluation metrics, as shown in Eqs. 12. In the equations, N_p is the predicted correct number of wheat spike, and N_g is the number of wheat spike true labels, and k is the number of pictures²³.

$$MAE = \frac{1}{k} \sum_{i=1}^k |N_p - N_g| \quad (12)$$

$$RMSE = \sqrt{\frac{1}{k} \sum_{i=1}^k |N_p - N_g|^2} \quad (13)$$

$$MSE = \frac{1}{k} \sum_{i=1}^k |N_p - N_g|^2 \quad (14)$$

$$R^2 = 1 - \frac{\sum_{i=1}^k (N_p - N_g)^2}{\sum_{i=1}^k (\overline{N_p} - \overline{N_g})^2} \quad (15)$$

Ablation experiments

We verify the effectiveness of these three methods through experiments and propose the optimal solution for each method. The results are shown in Table 2, which shows the effect of using different improvements. Baseline is the YOLOX and Baseline + IM + ASL + REP is our final model. It can be seen from the table that each method alone can increase the mAP of the model, which demonstrates the effectiveness of the three methods. Among them, the improvement effect brought by RepLKNet is the most obvious, and the mAP is 5.63% higher than the original YOLOX. At the same time, Recall and F1 increased by 11.14% and 9.48%, respectively. This demonstrates that the larger effective receptive field and shape information brought by the backbone significantly improve the performance of the model.

At the same time, we show the results of these three fusion methods. In the fusion experiment, we found from the results that the combination method gives better results. The combination of improvement points improves the performance of the original model. Similarly, RepLKNet plays an important role in the model, which makes the implicit decoupling of the detection head and the asymmetric loss more effective. The last three improvements produce the best results. The model mAP reaches 81.54%, which is 9.05% higher than the original model. The recall rate and F1 are increased by 15.07% and 13.20%, respectively.

To demonstrate the feature extraction capability of RepLKNet, we visualized the effective receptive fields of RepLKNet and the original backbone. Several works have demonstrated the importance of a large effective receptive field for detection^{36,38}. To examine the effective receptive field, we randomly selected 100 images from the dataset. We then conducted separate tests on both the original backbone and RepLKNet using these images. The size of the effective receptive field was determined by calculating the area of the receptive field per pixel unit. Figure 8 shows the visualization of the effective receptive field. Colors closer to red in the figure indicate larger gradient values, and colors closer to blue indicate smaller gradient values. The larger the hot spot region indicates a larger effective receptive field. Figure 9 is a set of images used to visually represent the effect of attention. Among them, Fig. 9(a) is the original image. Figure 9(b) shows the visualization results of the experiment using the baseline model YOLOX. Figure 9(c) presents the visualization results of the experiment using the model proposed in this paper. Figures 9(d) to Figs. 9(h) are used to indicate the data collection locations for the original image examples. From the visualization results in the figure, we can observe that the hotspot area in Fig. 9(c) is larger and denser compared to Fig. 9(b), indicating that the improved network has a better effective receptive field.

Finally, in order to compare the detection capabilities before and after the improvement, we visualize the heat map of the model. As shown in Fig. 8, we use Class Activation Mapping (CAM) to demonstrate the difference in detection performance^{40,41}. The different color heat map features in the figure reveal the attention of the model. In the figure, the red area has the greatest impact on the model, because the color changes from red to yellow and finally to blue, indicating that the impact is gradually weakening. It can be clearly seen from the figure that

Model	IM	ASL	REP	mAP	Precision	Recall	F1
YOLOX(Baseline)	-	-	-	72.49%	88.04%	56.25%	64.71%
Baseline + IM	√	-	-	74.27%	88.79%	56.82%	65.65%
Baseline + ASL	-	√	-	74.17%	88.50%	56.58%	64.79%
Baseline + REP	-	-	√	78.12%	88.59%	67.39%	74.19%
Baseline + IM + REP	√	-	√	80.17%	89.14%	69.02%	76.03%
Baseline + ASL + REP	-	√	√	80.06%	89.05%	67.90%	74.76%
Baseline + IM + ASL + REP(our)	√	√	√	81.54%	89.25%	71.32%	77.91%

Table 2. Results for each improvement point used respectively .REP, RepLKNet; IM, implicitly decoupled detection head; ASL, Asymmetric Loss.

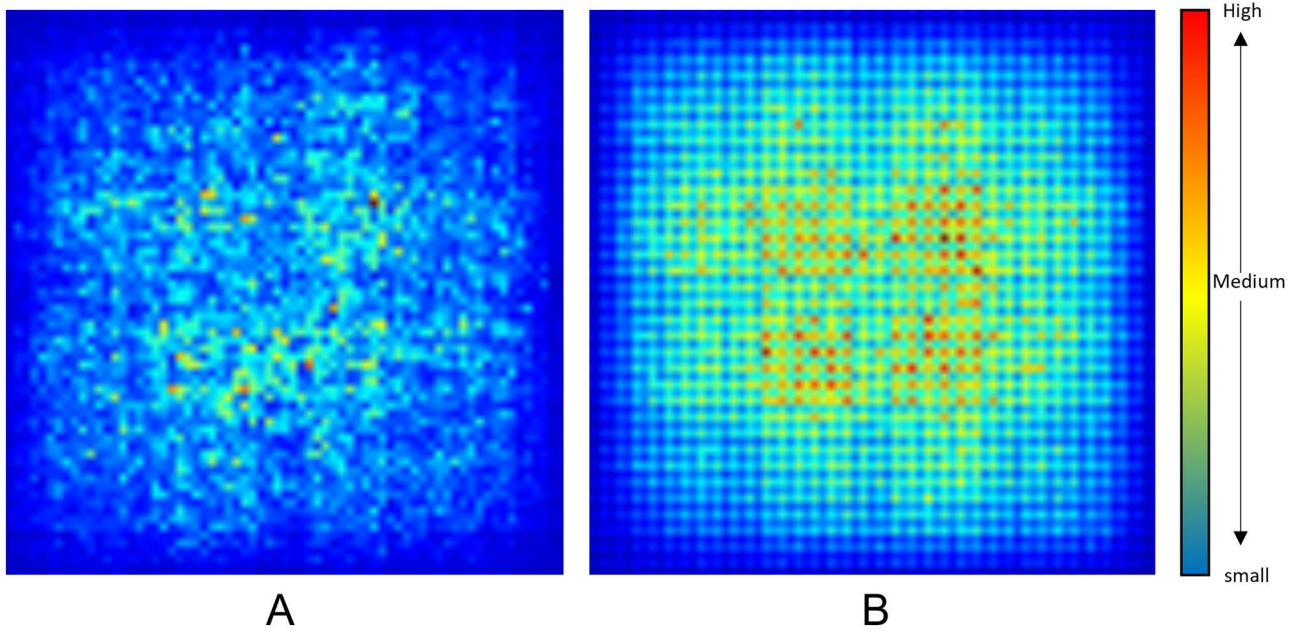


Fig. 8. Effective receptive field visualization of YOLOX backbone and RepLKNet backbone: (a) YOLOX backbone; (b) RepLKNet backbone.

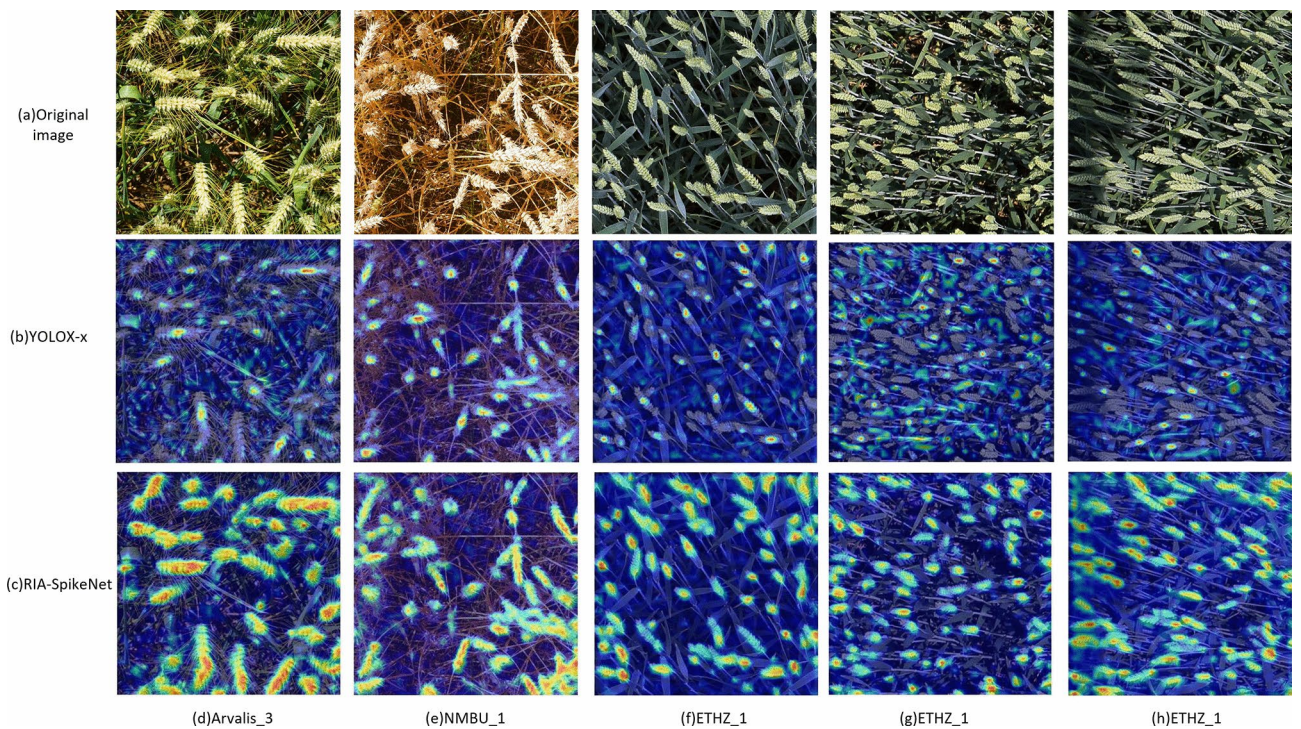


Fig. 9. CAM comparison between the improved model and YOLOX. (d): French wheat in the Ripening stage; (e): Norwegian wheat in the Filling stage; (f), (g), (h): Swiss wheat in the Filling stage.

our model can better detect wheat ears. The backgrounds of these images have similar color, but our network still accurately finds the wheat ears, which illustrates the role of the implicit model and improves the accuracy of the model. Moreover, it can be seen that our network has fewer false detections of wheat ears, which means that the number of negative samples has decreased, indicating that the loss function can control the balance of samples very well. At the same time, it can be found that the wheat ear features detected by our network almost cover the phenotype of the wheat ear, which indicates that the network has a stronger ability to extract phenotype features

Model	mAP	Precision	Recall	F1
YOLOv4[2019]	59.80%	65.70%	57.70%	57.46%
YOLOv5[2020]	68.90%	70.00%	65.90%	65.28%
YOLOR[2021]	78.80%	85.90%	70.10%	76.47%
YOLOX[2021]	72.49%	88.04%	56.25%	64.71%
GWC_YOLOv5[2021]	70.20%	75.60%	63.39%	64.87%
YOLOv6[2022]	76.60%	74.70%	70.00%	72.30%
YOLOv7[2022]	75.00%	77.50%	70.60%	73.39%
YOLOv8[2023]	79.94%	88.02%	71.32%	77.55%
YOLOv9m[2024]	85.12%	84.91%	78.70%	81.69%
YOLOv10x[2024]	86.30%	86.60%	79.60%	82.95%
RIA-SpikeNet (our)	81.54 %	89.25 %	71.35 %	77.91 %

Table 3. The comparison results with other detection models

	Model	RMSE	R ² (%)	MAE	MSE
Ablation experiments	YOLOX	11.21	76.44	6.03	125.66
	REP	10.96	77.47	5.54	120.12
	IM	12.68	69.84	6.41	160.78
	ASL	11.72	74.26	6.57	137.35
	REP + ASL	9.88	81.69	6.06	97.61
	RIA-SpikeNet(our)	7.19	90.29	5.25	51.69
	YOLOv4	16.86	44.60	8.93	284.25
Comparison experiment	YOLOv5	13.27	66.05	6.88	176.09
	YOLOR	8.21	87.14	6.14	67.40
	GWC_YOLOv5	10.90	76.99	6.76	118.81
	YOLOv6	8.60	86.10	6.02	66.42
	YOLOv7	11.59	74.22	6.72	134.32
	YOLOv8	8.15	87.53	5.53	65.85
	RIA-SpikeNet(our)	7.19	90.29	5.25	51.69

Table 4. The result of counting comparison between improved points REP, RepLKNNet; IM, implicitly decoupled detection head; ASL, Asymmetric Loss.

after improvement, which is the role of the RepLKNNet backbone, which solves the problem of complex texture and difficult detection.

Comparison experiments

We compare our model with classical and current advanced object detection models. The results of the comparative experiment are shown in Table 3. The compared networks include the classical YOLOv4 and YOLOv5, and the latest proposed models including YOLOR, YOLOv6, YOLOv7, and YOLOv8^{42–44}. From the table, we can see that our model achieves the highest mAP of 81.54. In particular, it is much higher than YOLOv4 and YOLOv5, which are 21.74% and 12.64% higher than YOLOv4 and YOLOv5, respectively. Our model also outperforms the current state-of-the-art models YOLOR, YOLOv6, and YOLOv8 in wheat spike detection. Our mAP is 2.74%, 4.11%, and 1.6% higher than YOLOR, YOLOv6, and YOLOv8, respectively. This demonstrates that our model has stronger comprehensive detection capabilities. In addition, our Precision, Recall, and F1 are 19.25%, 5.44%, and 12.62% higher than YOLOv5, respectively. At the same time, our Precision and F1 are even better, which are 14.54% and 5.60% higher than YOLOv6, respectively. Finally, we also compare with GWC_YOLOv5, which is one of the winning models of the Global Wheat Challenge 2021. We conduct experiments using the GWC_YOLOv5 model and the same experimental parameters as other experiments. As can be seen from Table 4, in this experiment, RIA-SpikeNet outperforms it in all metrics, with mAP and F1 being 11.34% and 13.03% higher than GWC_YOLOv5, respectively. The above comparison results fully demonstrate that our model has higher detection accuracy.

We carefully selected YOLOv9m and YOLOv10x for comparison based on their model sizes and detection performance, as these two variants exhibit remarkable capabilities. Although our proposed RIA-SpikeNet model lags slightly behind in several metrics, it demonstrates exceptional results in terms of Precision. While YOLOv9m and YOLOv10x exhibit higher overall mAP scores compared to our RIA-SpikeNet, it's important to note that RIA-SpikeNet demonstrates notable advantages in specific aspects of performance, highlighting the unique strengths of our algorithm. One key advantage of RIA-SpikeNet lies in its exceptional precision score of 89.25%, significantly outperforming both YOLOv9m and YOLOv10x. This high precision indicates that our

model is more accurate in classifying true wheat spikes, resulting in fewer false positives. In applications such as yield estimation, where the accuracy of spike counts directly impacts the final estimate, this high precision is a significant advantage. Additionally, while RIA-SpikeNet's recall score of 71.35% may seem lower than the other two models, it's worth noting that recall measures the ability to find all relevant objects, and a high recall can sometimes come at the cost of increased false positives. By maintaining a high precision while achieving a reasonable recall, RIA-SpikeNet strikes a balance that may be more desirable in certain scenarios, especially when reducing false positives is critical. Furthermore, the F1 score, which is a harmonic mean of precision and recall, provides a more holistic view of model performance. While RIA-SpikeNet's F1 score of 77.91% is slightly lower than the other models, it underscores the trade-off between precision and recall that our model achieves. In practice, this balance can be tailored to the specific requirements of the task, such as emphasizing precision over recall for applications where false positives are costly.

Count evaluation

Wheat spike counting is an important task, and more accurate counting is important for wheat yield estimation and breeding. We measured the counting performance of RIA-SpikeNet using a test set evaluation. As with the detection experiments, we conducted ablation experiments with improved points and calculated RMSE and R^2 for predicted and true labels, respectively. As can be seen from the Table 4, our model all have different degrees of improvement for counting performance. The final model achieves the optimal results with RMSE reduced by 4.02, and R^2 improved by 13.85%, which finally reached 90.29%. We also compared the effectiveness of counting with other detection models. From Table 4, we observe that our model is leading in terms of counting. The RMSE of our model is lower by 0.96 and R^2 is higher by 2.76% than YOLOv8. This demonstrates that our model has leading counting performance and can complete counting tasks very well.

To show the performance of model counting more clearly, we also plotted the relationship between the real labels and the predicted results. In Fig. 10, the X-axis represents the number of wheat tags and the Y-axis represents the number of wheat spikes predicted by the model. The red line is the straight line $y=x$, which indicates that all predictions are correct. The green line is the curve of linear fit of the predicted results. As can be seen from the figure, the prediction results of our model are all concentrated near $y=x$, and the fitted curve is closer to $y=x$, which demonstrates that our model has better counting performance.

Discussion on Complex scene detection

To demonstrate the advantages of RIA-SpikeNet in complex scenes, 155 images of complex environments were selected in our dataset for evaluation. These environments include low lighting conditions, similar colors, shadows, overlapping objects, and complex backgrounds. The number of wheat spikes in these images ranges from 80–120, which is larger than the average of 43 wheat spikes per single image. This indicates a highly dense wheat environment that is difficult to detect. Additionally, the number of images in each category within these pictures is only 20–100, which is less than the average of 141 images per category. This makes it difficult to train models on categories with small wheat spikes effectively. As can be seen in Table 5, RIA-SpikeNet has the highest mAP, Precision, Recall and F1 with other detection models. Among them, the mAP is 4.07% and 4.26% higher than YOLOR and YOLOv8, respectively, which is more prominent than the detection effect on the test set. Therefore, it shows that our model is stronger in complex scenes and overcomes the problems caused by complex scenes.

We also show the visualization of some models' detection results in complex scenarios. As shown in Fig. 11, the red boxes represent the correctly detected wheat spikes, and the yellow boxes represent the missed spikes. It can be observed from the figure that our model has the best detection performance in several scenes, especially in the dark scene and the complex background, it can be seen that our model only has a few missed detections. These results demonstrate the effectiveness of our model in addressing the challenging problem of wheat spike detection in various scenarios.

In order to better see the detection effect in complex scenarios, we count the detection results of these images, as shown in Table 6. From the table, we can see that the detection error of RIA-SpikeNet is the smallest, which is 30, 27, 26, and 4 in several environments. This proves that RIA-SpikeNet has better performance in complex environments.

Discussion and conclusion

Wheat spike detection and counting is of great significance and can be used for various purposes, including visual object detection, wheat yield estimation, crop management, and plant breeding. However, complex field scenes pose challenges to this task. In this paper, we present the RIA-SpikeNet network for wheat spike detection and counting in complex scenes. Firstly, we propose an implicit decoupling detection head. We add an implicit model to the decoupling detection header. The addition of the implicit model can enrich the feature information and refine it. Therefore, the implicit decoupled detection head can improve the identification accuracy of wheat spike in complex environment and solve the problem of similarity between wheat spike and background color gamut. We use asymmetric loss and emphasize the learning of samples by it adjusting the gradient of positive and negative sample loss. One way is to use decoupled attention for positive and negative samples by setting γ_- and γ_+ as the parameters for positive and negative attention, respectively. The second way is to add a hard threshold m to the loss of negative samples, which is used to reduce the gradient generated by the negative samples that are easily detected in the negative samples. Setting m is not only to reduce the number of negative samples, but also to enhance the learning of difficult samples. The learning of wheat ears is enhanced by adjusting the loss weights of positive and negative samples and the weights of difficult samples to solve the positive and negative sample imbalance and difficult sample learning problems in complex environments. By using RepLKNet as the backbone, the model has a larger effective receptive field and shape information. The larger effective receptive

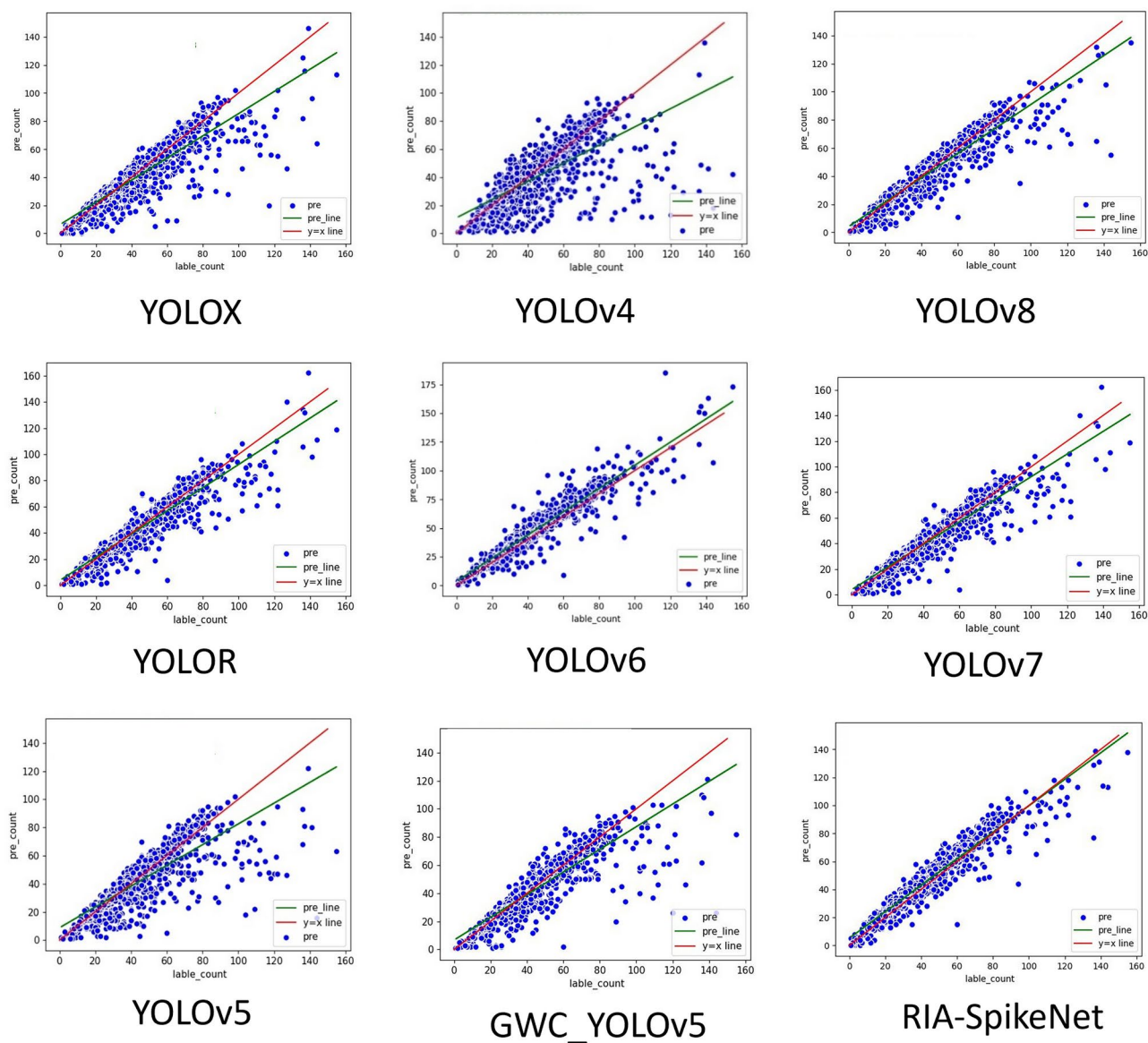


Fig. 10. Graph of the relationship between true labels and predicted results.

Model	mAP	Precision	Recall	F1
YOLOv4	38.00%	51.60%	37.70%	37.95%
YOLOv5	54.30%	63.60%	49.10%	49.85%
YOLOR	62.20%	74.90%	58.30%	64.70%
YOLOX	48.17%	62.16%	41.36%	41.10%
GWC_YOLOv5	55.90%	72.50%	49.00%	54.78%
YOLOv6	55.80%	63.30%	54.00%	58.30%
YOLOv7	58.30%	68.60%	55.50%	57.65%
YOLOv8	62.01%	75.22%	58.58%	64.06%
RIA-SpikeNet(our)	66.27%	75.95%	59.27%	65.45%

Table 5. Comparison results with other detection models on 155 images.

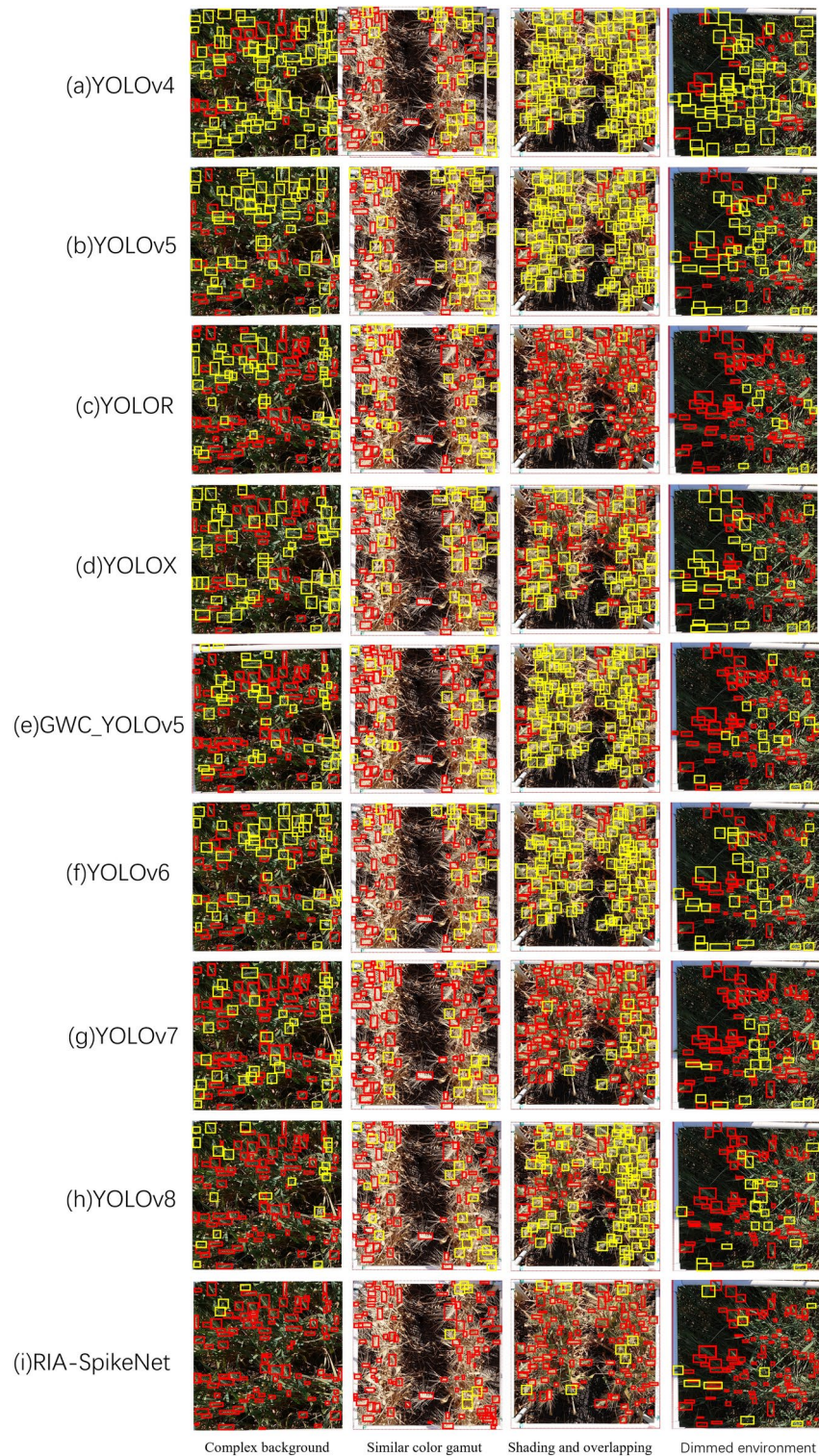


Fig. 11. Comparison of detection visualization in different environments.

field can capture more global feature information, prevent useful information from being missed, and enhance the information extraction of wheat spikes and background. More shape information can better identify wheat spikes from the background. The overlap and occlusion of wheat spikes make the wheat texture features more complex, which makes it more difficult to detect wheat spikes. Our network has more shape information and can better detect wheat spikes in complex textures. Enhancing the effective receptive field and shape information can better detect the wheat spikes in obscured and overlapping environments. The method achieved 81.54% mAP and 78.4 F1 score in the wheat spike detection experiment. In the wheat spike counting experiment, the RMSE

Model	Environments	Red boxes	Yellow boxes	All boxes	Ground boxes
YOLOv4	Dim environment	23	80	103	121
	Occlusion and overlap	5	121	126	144
	Similar gamut	16	83	99	122
	Complex background	22	72	94	95
YOLOv5	Dim environment	54	49	103	121
	Occlusion and overlap	16	110	126	144
	Similar gamut	47	52	99	122
	Complex background	39	55	94	95
YOLOvR	Dim environment	84	19	103	121
	Occlusion and overlap	112	14	126	144
	Similar gamut	61	38	99	122
	Complex background	63	31	94	95
YOLOvX	Dim environment	59	44	103	121
	Occlusion and overlap	63	63	126	144
	Similar gamut	54	45	99	122
	Complex background	41	53	94	95
GWC_YOLOv5	Dim environment	78	25	103	121
	Occlusion and overlap	24	102	126	144
	Similar gamut	63	36	99	122
	Complex background	80	14	94	95
YOLOv6	Dim environment	62	41	103	121
	Occlusion and overlap	31	95	126	144
	Similar gamut	63	36	99	122
	Complex background	52	42	94	95
YOLOv7	Dim environment	81	22	103	121
	Occlusion and overlap	106	20	126	144
	Similar gamut	78	21	99	122
	Complex background	73	21	94	95
YOLOv8	Dim environment	76	27	103	121
	Occlusion and overlap	51	75	126	144
	Similar gamut	63	36	99	122
	Complex background	78	16	94	95
RIA-SpikeNet(our)	Dim environment	94	9	103	121
	Occlusion and overlap	118	8	126	144
	Similar gamut	92	7	99	122
	Complex background	91	3	94	95

Table 6. Complex environment test results

of the method was less than 7.19 and the R^2 reached 90.29%. Therefore, this study will be helpful for wheat spike detection and counting in complex scenarios, and also provide a counting reference for wheat yield estimation and wheat breeding.

Considering the application requirements of the model, future improvements will not only focus on improving accuracy, but also take into account the practical requirements of production practices in order to better complete the detection and counting tasks. Below are the limitations of this work and potential directions for future research:

Limitations

1. To ensure lightweight development, the model was not deployed on devices for testing; all computations were carried out on a computer.
2. The dataset requires further refinement, as certain categories have a limited number of samples.
3. The methods used for dataset processing are relatively limited.

Future directions

1. Exploring more efficient and lightweight network architectures: Developing more efficient and lightweight network architectures that can be deployed on embedded devices or mobile phones, enabling real-time wheat spike detection and counting in the field.

- Dataset improvement: Increase the number of images in underrepresented categories and add more categories, including the detection of damaged wheat.
- Investigating more robust data augmentation strategies: Developing more robust data augmentation strategies that can effectively address various challenges in real-world scenarios, such as complex backgrounds, occlusions, and scale variations.

By addressing these research directions, we can further improve the performance and applicability of wheat spike detection and counting models, making them more useful for practical applications in agriculture.

Data Availability

The data and reproduction code link of the model proposed in this paper is as follows: <https://github.com/mzy-m/Spike-YOLO1.git>.

Received: 18 April 2024; Accepted: 7 October 2024

Published online: 15 October 2024

References

- Liu, C., Wang, K., Lu, H. & Cao, Z. Dynamic Color Transform Networks for Wheat Head Detection. *Plant Phenomics* **2022**, 2022/9818452 (2022).
- Han, F. & Li, J. Wheat Heads Detection via Yolov5 with Weighted Coordinate Attention. in *2022 7th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA)* 300–306 (IEEE, Chengdu, China, 2022). <https://doi.org/10.1109/ICCCBDA55098.2022.9778925>.
- Curtis, T. & Halford, N. G. Food security: the challenge of increasing wheat yield and the importance of not compromising food safety. *Ann. Appl. Biol.* **164**, 354–372 (2014).
- Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327**, 818–822 (2010).
- Khaki, S., Safaei, N., Pham, H. & Wang, L. WheatNet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. *Neurocomputing* **489**, 78–89 (2022).
- Schauberger, B., Ben-Ari, T., Makowski, D. et al. Yield trends, variability and stagnation analysis of major crops in France over more than a century. *Sci Rep.* **8**(1), 1–12, (2018).
- Ferrante, A., Cartelle, J., Savin, R. & Slafer, G. A. Yield determination, interplay between major components and yield stability in a traditional and a contemporary wheat across a wide range of environments. *Field Crops Res.* **203**, 114–127 (2017).
- Xu, X. et al. Wheat ear counting using K-means clustering segmentation and convolutional neural network. *Plant Methods* **16**, 1–13 (2020).
- Laabassi, K., Belarbi, M. A., Mahmoudi, S., Mahmoudi, S. A. & Ferhat, K. Wheat varieties identification based on a deep learning approach. *J. Saudi Soc. Agric. Sci.* **20**, 281–289 (2021).
- Hasan, M. M., Chopin, J. P., Laga, H. & Miklavcic, S. J. Detection and analysis of wheat spikes using Convolutional Neural Networks. *Plant Methods* **14**, 100 (2018).
- Zhaosheng, Y., Tao, L., Tianle, Y., Chengxin, J. & Chengming, S. Rapid Detection of Wheat Ears in Orthophotos From Unmanned Aerial Vehicles in Fields Based on YOLOX. *Front. Plant Sci.* **13**, 851245 (2022).
- David, E. et al. Global Wheat Head Detection (GWHDD) Dataset: A Large and Diverse Dataset of High-Resolution RGB-Labelled Images to Develop and Benchmark Wheat Head Detection Methods. *Plant Phenomics* **2020**, 1–12 (2020).
- David, E. et al. Global Wheat Head Dataset 2021: more diversity to improve the benchmarking of wheat head localization methods. **11**.
- Fernandez-Gallego, J. A., Kefauver, S. C., Gutiérrez, N. A., Nieto-Taladriz, M. T. & Araus, J. L. Wheat ear counting in-field conditions: high throughput and low-cost approach using RGB images. *Plant Methods* **14**, 22 (2018).
- Qiongyan, L., Cai, J., Berger, B., Okamoto, M. & Miklavcic, S. J. Detecting spikes of wheat plants using neural networks with Laws texture energy. *Plant Methods* **13**, 83 (2017).
- Hasan, M. M., Chopin, J. P., Laga, H. & Miklavcic, S. J. Detection and analysis of wheat spikes using Convolutional Neural Networks. *Plant Methods* **14**, 100 (2018).
- Yang, B., Gao, Z., Gao, Y. & Zhu, Y. Rapid Detection and Counting of Wheat Ears in the Field Using YOLOv4 with Attention Module. *Agronomy* **11**, 1202 (2021).
- Wang, Y., Qin, Y. & Cui, J. Occlusion Robust Wheat Ear Counting Algorithm Based on Deep Learning. *Front. Plant Sci.* **12**, 645899 (2021).
- Woo, S., Park, J., Lee, J.-Y. & Kweon, I. S. CBAM: Convolutional Block Attention Module. <https://doi.org/10.48550/arXiv.1807.06521> (2018).
- Sun, J. et al. Wheat head counting in the wild by an augmented feature pyramid networks-based convolutional neural network. *Comput. Electron. Agric.* **193**, 106705 (2022).
- Guo, C., Fan, B., Zhang, Q. et al. Augfpn: Improving multi-scale feature learning for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 12595–12604, (2020).
- Khan, S. D., Basalamah, S. & Naseer, A. Classification of plant diseases in images using dense-inception architecture with attention modules. *Multimed. Tools Appl.* <https://doi.org/10.1007/s11042-024-19860-y> (2024).
- Wen, C. et al. Wheat Spike Detection and Counting in the Field Based on SpikeRetinaNet. *Front. Plant Sci.* **13**, 821717 (2022).
- Zang, H. et al. Detection method of wheat spike improved YOLOv5s based on the attention mechanism. *Front. Plant Sci.* **13**, 993244 (2022).
- Zaji, A. et al. AutoOLA: Automatic object level augmentation for wheat spikes counting. *Comput. Electron. Agric.* **205**, 107623 (2023).
- Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection[J]. arxiv preprint arxiv:2004.10934, (2020).
- Qing, S. et al. Improved YOLO-FastestV2 wheat spike detection model based on a multi-stage attention mechanism with a LightFPN detection head. *Front. Plant Sci.* **15**, 1411510 (2024).
- Meng, X. et al. YOLOv7-MA: Improved YOLOv7-Based Wheat Head Detection and Counting. *Remote Sens.* **15**, 3770 (2023).
- Khan, S. D., Alarabi, L. & Basalamah, S. A Unified Deep Learning Framework of Multi-scale Detectors for Geo-spatial Object Detection in High-Resolution Satellite Images. *Arab. J. Sci. Eng.* **47**, 9489–9504 (2022).
- Zheng, G., Songtao, L., Feng, W. et al. YOLOX: Exceeding YOLO series in 2021[J]. arxiv preprint arxiv:2107.08430, (2021).
- Farhadi, A., Redmon, J. Yolov3: An incremental improvement[C]//Computer vision and pattern recognition. Berlin/Heidelberg, Germany: Springer, **1804**, 1–6, (2018).
- Tian, Z., Shen, C., Chen, H. & He, T. FCOS: A Simple and Strong Anchor-Free Object Detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1922–1933 (2022).

33. Hendrycks, D. & Gimpel, K. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. (2016).
34. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. in *Proceedings of the 32nd International Conference on Machine Learning* 448–456 (PMLR, 2015).
35. Hosang, J., Benenson, R., Schiele, B. Learning non-maximum suppression[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 4507–4515. (2017).
36. Wang, C.Y., Yeh, I.H., Liao, H.Y.M. You only learn one representation: Unified network for multiple tasks[J]. arxiv preprint arxiv:2105.04206, (2021).
37. Ding, X., Zhang, X., Han, J. et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11963–11975. (2022).
38. Luo, W., Li, Y., Urtasun, R. & Zemel, R. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks.
39. Wen, C. et al. Pest-YOLO: A model for large-scale multi-class dense and tiny pest detection and counting. *Front. Plant Sci.* **13**, (2022).
40. Namburu, S. M., Tu, H., Luo, J. & Pattipati, K. R. Experiments on Supervised Learning Algorithms for Text Categorization. in *2005 IEEE Aerospace Conference* 1–8 (2005). <https://doi.org/10.1109/AERO.2005.1559612>.
41. Zhou, Q. et al. A wheat spike detection method based on Transformer[J]. *Frontiers in Plant Science* **13**, 1023924 (2022).
42. Jocher Glenn. YOLOv8. <https://github.com/ultralytics/ultralytics/tree/main>, 2023.
43. Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning what you want to learn using programmable gradient information[J]. arxiv preprint: arxiv:2402.13616, 2024.
44. Wang A, Chen H, Liu L, et al. YOLOv10: Real-time end-to-end object detection[J]. arxiv preprint: arxiv:2405.14458, 2024.

Acknowledgements

We want to thank "Changchun Computing Center of Huawei" and "Eco-Innovation Center" for providing inclusive computing power and technical support of MindSpore during the completion of this paper.

Author contributions

W: conceptualization, methodology, formal analysis, investigation, writing—review and editing. M: conceptualization, methodology, software, validation, formal analysis, investigation, Data curation, writing—original draft, writing—review and editing, visualization. CHR: formal analysis, writing—review and editing. ZT: investigation, data curation, formal analysis. ZL: investigation, data curation, formal analysis. R: supervision, project management. S: formal analysis, supervision, project management. analysis, supervision, project management. CHB: formal analysis, supervision, project management. G: formal analysis, survey, Data curation, writing—review and editing. We want to thank "Changchun Computing Center" and "Eco-Innovation Center" for providing inclusive computing power and technical support of MindSpore during the completion of this paper.

Funding

The research was funded by the Industrial Technology and Development Project of Development and Reform Commission of Jilin Province (No.2021C044–8, 2023C030–3), Key Laboratory open fund of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources (No.KF-2021–06–067), The research and planning project of Jilin Provincial Department of Education (No.JJKH20240441HT, JJKH20220376SK), Jilin Provincial Science and Technology Development Plan Project (No.20210203013SF) and the National Natural Science Foundation of China (Key Program)(No.U19A2061).

Declarations

Competing interests

The 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.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-75523-w>.

Correspondence and requests for materials should be addressed to C.W. or W.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024