agriculture

Article

WH-DETR: An Efficient Network Architecture for Wheat Spike Detection in Complex Backgrounds

Zhenlin Yang [1], Wanhong Yang [1], Jizheng Yi [2,*] and Rong Liu [2]

[1] College of Computer Science and Mathematics, Central South University of Forestry and Technology, Changsha 410004, China; 20212692@csuft.edu.cn (Z.Y.); 20211685@csuft.edu.cn (W.Y.)
[2] College of Advanced Interdisciplinary Studies, Central South University of Forestry and Technology, Changsha 410004, China; t20080437@csuft.edu.cn
* Correspondence: t20152279@csuft.edu.cn

Abstract: Wheat spike detection is crucial for estimating wheat yields and has a significant impact on the modernization of wheat cultivation and the advancement of precision agriculture. This study explores the application of the DETR (Detection Transformer) architecture in wheat spike detection, introducing a new perspective to this task. We propose a high-precision end-to-end network named WH-DETR, which is based on an enhanced RT-DETR architecture. Initially, we employ data augmentation techniques such as image rotation, scaling, and random occlusion on the GWHD2021 dataset to improve the model's generalization across various scenarios. A lightweight feature pyramid, GS-BiFPN, is implemented in the network's neck section to effectively extract the multi-scale features of wheat spikes in complex environments, such as those with occlusions, overlaps, and extreme lighting conditions. Additionally, the introduction of GSConv enhances the network precision while reducing the computational costs, thereby controlling the detection speed. Furthermore, the EIoU metric is integrated into the loss function, refined to better focus on partially occluded or overlapping spikes. The testing results on the dataset demonstrate that this method achieves an Average Precision (AP) of 95.7%, surpassing current state-of-the-art object detection methods in both precision and speed. These findings confirm that our approach more closely meets the practical requirements for wheat spike detection compared to existing methods.

Keywords: deep learning; detection transformer; feature pyramid; wheat spike detection; agriculture

check for updates

Citation: Yang, Z.; Yang, W.; Yi, J.; Liu, R. WH-DETR: An Efficient Network Architecture for Wheat Spike Detection in Complex Backgrounds. *Agriculture* **2024**, *14*, 961. https://doi.org/10.3390/agriculture14060961

Academic Editor: Francesco Marinello

Received: 8 May 2024
Revised: 16 June 2024
Accepted: 17 June 2024
Published: 19 June 2024

1. Introduction

Wheat is regarded as one of the "big three" cereal crops due to its extensive global cultivation range [1]. Ten thousand years ago, hunter-gatherers began cultivating wild emmer wheat [2]. Today, wheat is one of the world's most crucial food sources, with an estimated production of 802.8 billion tons projected for the 2023–2024 season [3]. Estimating wheat yield has always been a significant issue in agriculture, profoundly influencing the modernization of wheat cultivation and the advancement of precision agriculture [4]. The accurate counting of wheat spikes is one of the key factors in efficient agricultural management and resource allocation. Therefore, the precise detection and identification of wheat spikes are crucial for implementing precision agricultural management, optimizing agricultural production, and enhancing crop yields.

Wheat spike detection has consistently been a significant challenge in the field of object detection, drawing global talent to the Global Wheat Detection competition [5] to devise solutions. Conventional methods for estimating wheat yield often involve lengthy processes and are susceptible to inaccuracies [6]. As computer vision technology has advanced, a growing number of researchers have started to utilize machine learning approaches in their studies. Pantazi et al. [7] utilized high-resolution, multi-layer soil data and satellite imagery to predict variations in wheat yield within fields. Franch et al. [8] assessed and predicted the winter wheat yields in the USA and Ukraine using the Difference Vegetation Index (DVI) derived from MODIS satellite data. Rocha and Dias [9] proposed a Radial Basis Function (RBF) interpolation model for the early-season yield prediction of durum wheat in Spain. Although classic machine learning techniques have achieved some success in predicting wheat yield, their performance heavily relies on cumbersome and error-prone feature engineering, particularly evident in complex scenarios [10]. The lack of robustness in these techniques primarily manifests in their severe dependence on manual feature extraction and hyperparameter tuning, resulting in a poor performance under uncontrolled environmental conditions such as field illumination, weather, and exposure. Therefore, without meticulous feature extraction under the supervision of domain experts, these methods often struggle to adapt to uncontrolled real-world application settings, affecting the accuracy and practicality of the models.

Leveraging deep learning, particularly object detection technology, has proven to be a potential method for addressing this issue [11]. Deep learning can autonomously extract and learn features from raw data, reducing the need for manual feature engineering. Misra et al. [12] combined digital image analysis and machine learning to develop SpikeSegNet for detecting and counting wheat spikes, achieving significant results. Chandra et al. [13] introduced a point-supervised active learning method for detecting wheat spikes. Hasan et al. [14] employed four types of R-CNN models to detect wheat spikes and assess wheat yields under different varieties and fertilizer treatments. Madec et al. [6] explored two distinct approaches using the Faster-RCNN network and TasselNet local count regression network to derive wheat spike density from high-resolution RGB images. Gong et al. [15] enhanced the YOLOv4 network by introducing a Dual Spatial Pyramid Pooling (SPP) network, proposing a high-accuracy and high-speed real-time detection method for wheat spikes. Sun et al. [16] proposed an improved wheat head counting network, WHCnet, employing an enhanced feature pyramid network (AugFPN) to address issues with poor wheat head detection. Ye et al. [17] introduced a real-time lightweight neural network named WheatLFANet for the efficient detection and counting of wheat heads, suitable for deployment on low-end devices. Yan et al. [18] developed a method for refining the scale of detection layers in a wheat spike detection network using the deep learning interpretation method GradCAM. Zhao et al. [19] introduced WheatNet for detecting wheat spikes from the filling to maturity stages. CNNs have a strong local perception ability, enabled by stacking multiple convolutional layers to expand the field of view and effectively capture local features in images. However, this also results in limited context capture, a drawback particularly evident in the complex scenarios of wheat spike detection. Wheat spike object detection still faces challenges such as overlap and crossing, occlusions and shadows, light transformations, changes in angle and scale, varietal differences, and growth environments [17], which hinder the performance improvement of wheat spike detection. The accuracy issues of the network, computational efficiency, and adaptability under different environmental conditions remain pressing concerns to be addressed.

Object detection algorithms are a crucial technology in the field of computer vision, primarily divided into two categories: one-stage and two-stage algorithms. Two-stage algorithms initially generate region proposals, then classify each region and regress its bounding box. R-CNN [20] marked the inception of two-stage algorithms, generating region proposals, extracting features for each region, and classifying them using an SVM classifier. Faster R-CNN [21] introduced the Region Proposal Network (RPN), which automatically generates high-quality region proposals, enhancing the speed and accuracy of object detection. One-stage algorithms predict object categories and bounding boxes directly on the image without the need for generating region proposals, thus, they are typically faster. YOLO [22] (You Only Look Once) introduced a revolutionary one-stage detection method, significantly simplifying the object detection process by treating it as a single regression problem, directly mapping from image pixels to bounding box coordinates and class probabilities. YOLOv4 [23] introduced innovations like Mosaic data augmentation and new anchor-free detection heads. YOLOv5 [24] proposed the

CSPNet structure to reduce computation and introduced automatic anchor adjustment. YOLOv7 [25] proposed the expanded Efficient Layer Aggregation Network (E-ELAN), enhancing the feature learning between different feature maps. Subsequent YOLOv8 [26] and other iterations of the YOLO algorithm have continuously innovated, leading to sustained improvements in network performance. Besides the YOLO series, SSD [27] (Single Shot MultiBox Detector) performs object detection on feature maps of different scales, balancing speed and accuracy. RetinaNet [28] addressed the issue of class imbalance in one-stage detectors, introducing Focal Loss to enhance the detection performance for hard-to-classify samples, significantly increasing the detection accuracy. CenterNet [29] transforms object detection into a keypoint detection task, directly predicting object centers and sizes to improve detection speed and accuracy. EfficientDet [30] combines EfficientNet with a compound scaling method, reducing parameters and computational costs while maintaining a high accuracy. The evolution of object detection algorithms from two stages to one stage has not only made significant progress in speed and efficiency, but has also continuously improved detection accuracy.

Recently, the Transformer architecture [31] has become prominent in computer vision, notably due to its success in natural language processing tasks. This architecture leverages a self-attention mechanism to analyze interdependencies within sequences, effectively gathering global contextual data and delivering outstanding results in pixel-level object detection. In 2020, Carion et al. [32] developed DETR, which applies the Transformer's encoder–decoder framework for direct object detection modeling, discarding the conventional reliance on anchor boxes and intricate post-processing actions like Non-Maximum Suppression (NMS), thereby reconceptualizing object detection into a direct set prediction task. Following this, Zhu et al. [33] introduced a Deformable DETR that incorporates offset attention sampling and deformable convolution to mitigate DETR's slow convergence and subpar performance with small objects. Subsequently, Meng et al. [34] enhanced DETR's design with a Conditional DETR that integrates a conditional mechanism to boost DETR's efficiency and effectiveness. Moreover, Lv et al. [35] unveiled RT-DETR, which includes an Efficient Hybrid Encoder and IoU-aware Query Selection, optimizing computational efficiency and accuracy for real-time detection. Despite these advancements, computational complexity and inference speed are ongoing challenges in real-time applications.

Within object detection networks, the Neck section serves as a critical intermediary between the Backbone, which extracts image features, and the Detection Head, which handles object classification and localization. It plays a pivotal role by refining and processing the features gathered by the Backbone. This refinement aids in precise feature fusion and enhances spatial contextual awareness, enriching the Detection Head with more distinct feature sets. The architecture of the Neck is crucial for boosting detection efficacy, as it influences feature quality and overall detection precision directly. Utilizing multi-scale feature pyramids enhances the Neck's ability to handle objects of different sizes and shapes effectively. By constructing feature maps at various scales, this approach gathers detailed and contextual data from the images. The established Feature Pyramid Network (FPN) [36] significantly boosts the detection of small and occluded objects by merging detailed low-level and semantic high-level information. Our research explores substituting standard convolutions in the Neck with GSConv [37], a lightweight convolution technology, and merging it with the weighted bi-directional feature pyramid (BiFPN) [30]. This integration forms a novel lightweight feature pyramid, GS-BiFPN, facilitating efficient and swift multi-scale feature fusion.

In object detection model training, loss functions are essential for measuring the discrepancies between predicted outcomes and actual results, which facilitates model refinement. Commonly used loss functions in machine learning, such as cross-entropy and mean squared error, are prevalent. Yet, these traditional forms may not sufficiently address the classification and localization errors within complex detection tasks. To overcome these limitations, advanced loss functions like Focal Loss [28] and IoU Loss [38] have been designed to adeptly manage the problems of scale variation and class imbalances. This research incorporates EIoU [39] and assesses the suitability and impact of tailored loss functions for the precise detection of wheat spikes.

Previous studies on wheat spike detection tasks have primarily focused on using traditional CNN-based object detection algorithms, such as Faster R-CNN and YOLO, for detecting and counting wheat spikes, enhancing detection accuracy and efficiency by adjusting network structures and parameters. Although Transformer-based methods have achieved significant success in other computer vision fields, they are relatively new in object detection, and there has been limited research on their application to wheat spike detection tasks. Inspired by the Swin Transformer architecture [40], Zhou et al. [41] proposed a Transformer-based wheat spike detection network, MW-Swin Transformer, marking the first effort to apply Transformers in the wheat detection field. Zhu et al. [42] developed three object detection methods using Transformer as the backbone to detect wheat spikes. This study represents another effort of using Transformer in wheat detection and the first application of the DETR series algorithms in this field. The goal of this research is to further investigate wheat spike detection using the DETR algorithm architecture as the research foundation, proposing a high-accuracy object detection network for wheat spikes. This aims to apply the DETR architecture to the field of wheat detection, offering a new approach to wheat detection and achieving a high accuracy while maintaining a good real-time performance. The contributions of this article are as follows:

- A lightweight feature pyramid, GS-BiFPN, is proposed in the Neck section of the neural network, which achieves multi-scale feature fusion for wheat spikes under complex conditions, addressing issues caused by the overlapping and occlusion of wheat spikes. Additionally, GSConv is comprehensively introduced to further reduce computational costs.
- EIoU is integrated into the original loss function, resulting in an improved loss function that increases focus on hard-to-detect portions of wheat spikes, such as those partially obscured or overlapped, while also accelerating the convergence of the model training process.
- Evaluation on a mixed dataset shows that our approach yielded promising results. The Average Precision (AP) of our system achieved 95.7%, exceeding the performance of existing leading object detection technologies. This performance confirms that our technique satisfies the stringent accuracy demands necessary for detecting wheat spikes in real-world farming environments.

2. Materials and Methods

2.1. Data Acquisition

This research utilizes the dataset from the Global Wheat Head Detection Challenge (GWHD2021), detailed on its official website: http://www.global-wheat.com (accessed on 9 August 2023). The 2021 edition of the GWHD dataset [43] includes an additional 1722 images from five countries, adding 81,553 wheat heads to the dataset. In total, the GWHD_2021 dataset features 6422 images and 275,187 wheat spikes contributed by 16 institutions from 12 countries/regions, as depicted in Figure 1. The image specifications include a resolution of 1024 × 1024 pixels, taken from heights of 1.8 m to 3 m. The cameras used vary in focal length from 10 mm to 50 mm and include different sensor sizes. The dataset includes diverse wheat genotypes from Europe, North America, Australia, and Asia, with planting densities ranging from low to high and row spacings from 12.5 cm to 30.5 cm. It also covers various soil and climatic conditions, including fertile loamy soil in France's Picardy and chalky clay in the Swiss highlands [44]. Using the GWHD2021 dataset promotes better model generalization. Additionally, the GWHD2021 dataset includes images capturing wheat spikes at various growth stages, including the late flowering, grain filling, grain filling maturity, and maturity stages. This diversity in growth stages ensures the robustness and adaptability of the trained models across different phenological phases. For this study, 5000 images containing 213,685 wheat heads were selected from the GWHD dataset. These were split into training and testing sets at an 80:20 ratio, with 4000 training images and 1000 testing images.

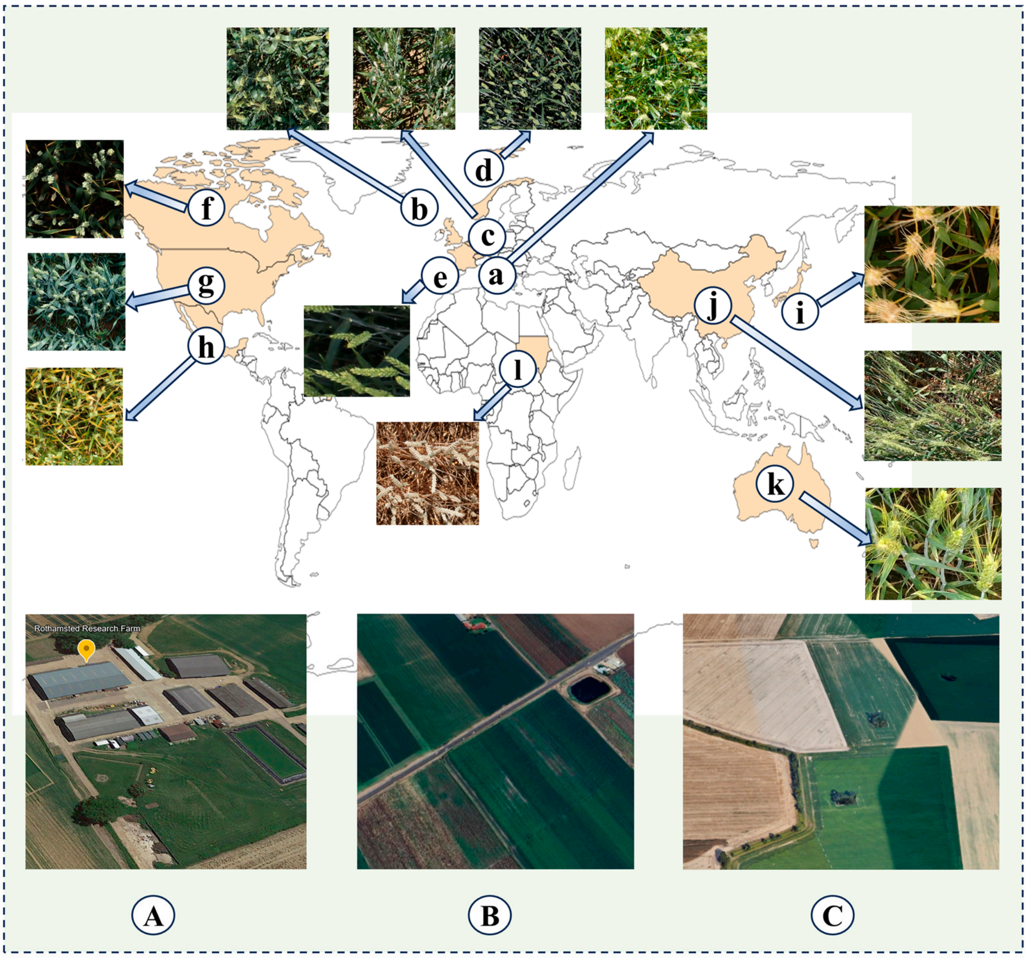

Figure 1. Distribution map of the Global Wheat Head Detection Challenge (GWHD2021) dataset. This dataset contains 6422 images and 275,187 wheat spikes, covering 16 institutions across 12 countries/regions. Schematics A–C depict actual agricultural field images from some of the dataset's collection locations. (**A**) Rothamsted Research Farm in the UK shows the row distribution of wheat spikes and the overall layout of the farm. (**B**) Gatton, Queensland, Australia, shows the wheat planting patterns adapted to the tropical climate. (**C**) Fields in Villiers le Bâcle, France, represent wheat cultivation under the temperate climate of Europe. Images (**a**–**l**) are actual field samples of wheat from various regions and environmental conditions, with specific countries/regions corresponding to the actual markings in the image. For example, (**a**) shows wheat cultivation in the mountainous conditions of the Swiss highlands. (**i**) Displays typical wheat from Japan. (**k**) Shows a typical dense distribution of wheat spikes in Australia.

2.2. Data Augmentation

Data augmentation plays a critical role in enhancing the robustness and generalizability of deep learning models, especially in the context of agricultural image analysis, where variations in lighting, orientation, and scale are common. In our research, we implemented various data augmentation strategies to artificially expand the diversity of the dataset, ensuring that our trained models performed exceptionally and maintained stable predictive capabilities when facing overfitting.

The augmentation techniques included geometric transformations such as random rotations (0° to 360°), horizontal and vertical flips, and random cropping. These transformations simulated the natural orientations and positions of wheat spikes in the field. Random rescaling was also applied to the images, simulating the effects of different camera distances and zoom levels, crucial for the varied camera focal lengths (10 mm to 50 mm) in the GWHD2021 dataset. To account for variations in lighting conditions, we introduced photometric enhancements such as brightness and contrast adjustments, as well as random shadows and highlights, creating a range of simulated lighting scenarios that might occur during data collection. Noise injection was another key aspect of our data augmentation strategy. This involved adding Gaussian noise to images, which helped the model become noise invariant, thereby stabilizing it against sensor noise or granularity, common under low-light conditions or with images from different sensor sizes in our dataset. Additionally, color enhancements were performed by altering the hue, saturation, and value (HSV) of the images, enabling the model to handle variations in soil and crop colors caused by different soil climate conditions and wheat genotypes in the dataset. Lastly, we employed random erasing by randomly deleting parts of images to introduce occlusions. This strategy was particularly effective for training the model to detect wheat spikes in the presence of obstacles, such as leaves or stems that might cover some spikes in the field. The augmented images were then merged into the training set to produce the final dataset, which included a balanced mix of original and augmented images, maintaining an 8:2 training to testing ratio. The effectiveness of the data augmentation techniques is shown in Figure 2. The augmented dataset comprised a total of 6750 images, with 5400 for training and 1350 for testing. The implementation of this comprehensive data augmentation strategy was expected to equip the model with the ability to reliably detect wheat spikes under various environmental conditions, reducing the model's sensitivity to variance that could lead to detection inaccuracies.

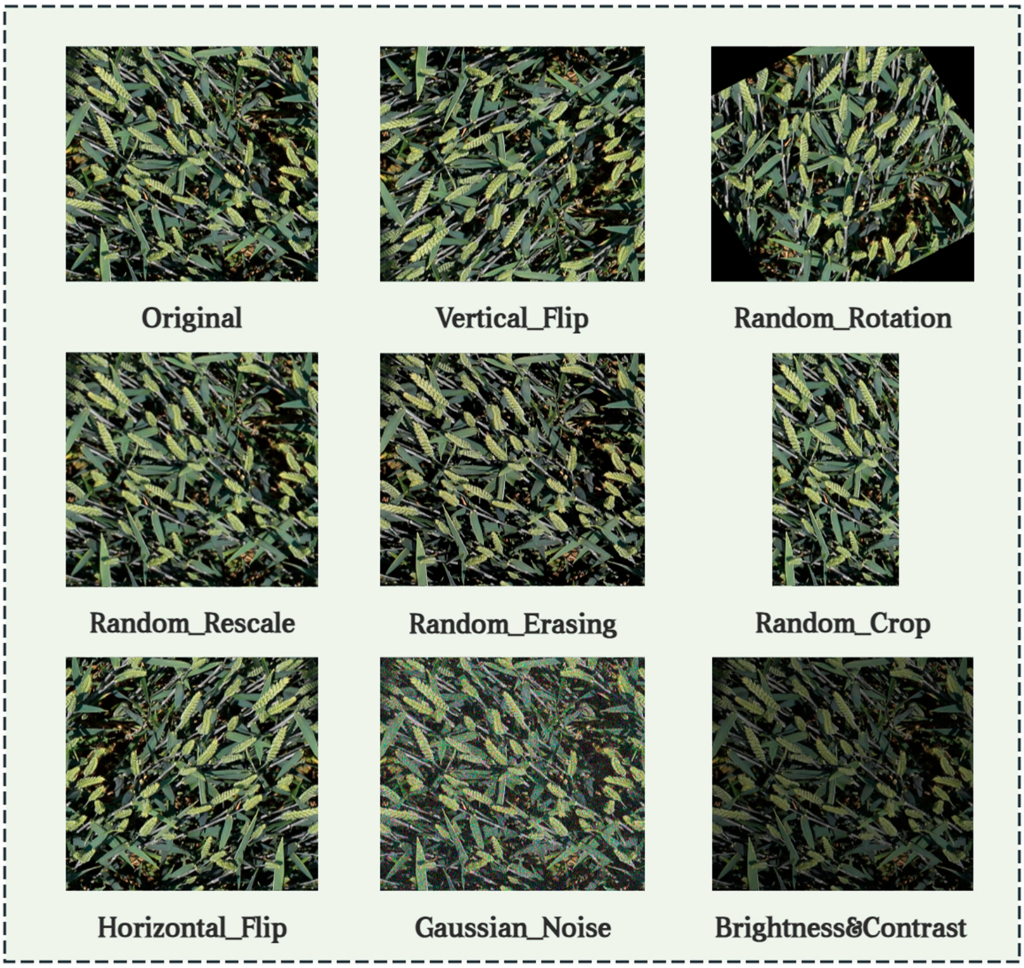

Figure 2. This image illustrates several data augmentation techniques applied to wheat spike images to simulate environmental variations. Techniques include the original, unmodified control; vertical

and horizontal flips for different orientations; random rotation for a 360° perspective; random rescale to simulate camera distance changes; random erasing and cropping to adjust for obstructions and positioning; and modifications to brightness, contrast, and noise to mimic natural lighting and enhance robustness.

2.3. Overall Architecture

In developing the WH-DETR model, we pinpointed several key challenges related to detecting wheat spikes within intricate agricultural settings, such as occlusions and overlapping, alongside demands for prompt detection capabilities. To tackle these issues, we adopted the RT-DETR network as our foundational architecture and implemented specific enhancements. Illustrated in Figures 3 and 4, we employed a lightweight feature pyramid network, GS-BiFPN, aimed at efficiently extracting multi-scale features with reduced computational demands. We further cut down on computational expenses by substituting standard convolutions with GSConv in the neck portion of the model. Additionally, we refined the loss function by incorporating EIoU loss, augmenting the model's proficiency in recognizing occluded and overlapping wheat spikes. Anticipated to boost the detection precision in complicated environments, these modifications also strived to uphold swift detection speeds to fulfill the prerequisites for real-time applications.

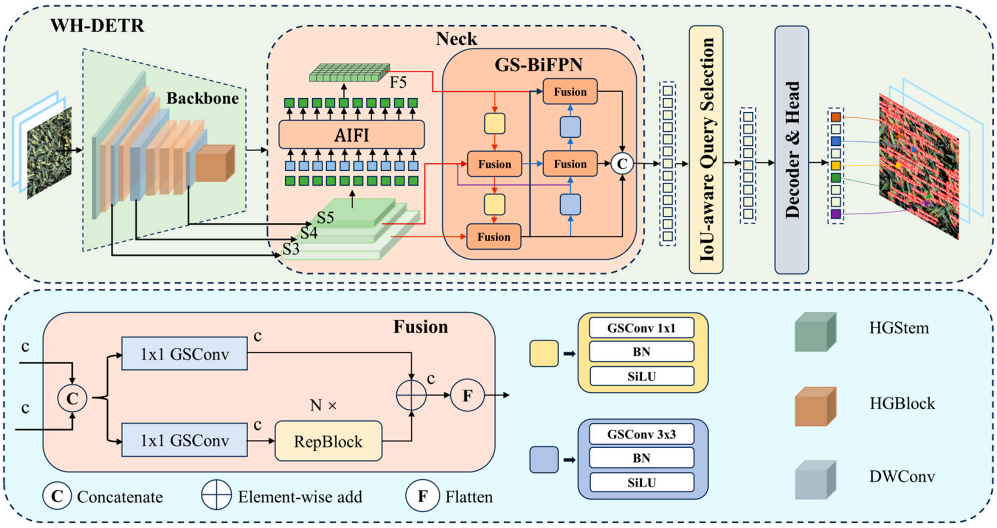

Figure 3. Introduction to WH-DETR. The model starts by harnessing features from the backbone's final three stages {S3, S4, and S5} for the encoder input. It features an efficient hybrid encoder that processes these multi-scale features into a coherent sequence of image features, facilitated by intra-scale feature interaction (AIFI) and the integration of GS-BiFPN. The system utilizes IoU-aware query selection to choose a predefined set of image features as initial queries for the decoder. Subsequently, the decoder, equipped with an auxiliary prediction head, methodically refines these queries to produce bounding boxes and confidence scores.

2.4. GS-BiFPN: A Strategy for Feature Fusion and Efficiency Optimization in Wheat Spike Detection

To improve wheat spike detection performance, especially in scenarios with frequent overlapping or occlusions, and considering the reasonable use of computational resources, this paper introduces a lightweight feature pyramid network architecture named GS-BiFPN. As shown in Figure 5, the proposed GS-BiFPN structure optimizes the bi-directional feature fusion path of the feature pyramid by directly connecting high-resolution P3 feature maps with lower-resolution P5, P6, and P7 feature maps. As shown in part (c) of the figure, GS-BiFPN strengthens the information flow through red connection lines, ensuring that detailed information is effectively transmitted even within multi-layered deep networks, enhancing the model's capability to detect wheat spikes. Part (d) of the figure, the GSConv component, demonstrates how we utilize depth-wise separable convolutions (DWConv), concatenation (Concat), and channel shuffling (shuffle) operations to reduce the computational load while maintaining the efficiency of feature fusion. These innovative designs fully exploit each level's feature map, from P3 to P7 as shown, improving the overall network's performance in wheat spike detection tasks. This strategy, based on the degradation phenomenon of detailed information propagation in deep networks, aims to mitigate its impact on wheat detection accuracy. Through this optimized information flow, the model's detection head can more directly utilize high-resolution features, becoming more sensitive to minute details of wheat spikes. Figure 6 reveals, in detail, how our proposed GS-BiFPN is implemented within the WHDETR model, where the optimized feature transfer mechanism enhances detailed information transfer while ensuring computational efficiency. The innovative GS-BiFPN component of our architecture is described in detail in Algorithm 1. This pseudocode explains how GS-BiFPN facilitates multi-scale feature integration within the network. This comprehensive design strategy ensures that the lightweight network model does not sacrifice detection accuracy.

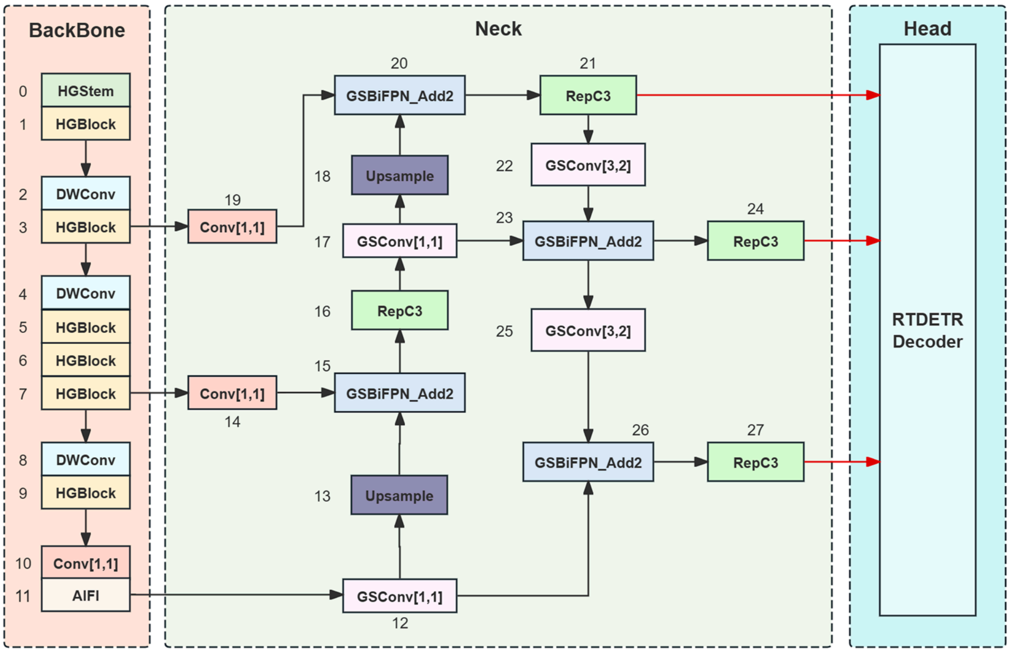

Figure 4. WH-DETR network structure.

Algorithm 1: Pseudocode of the GS-BiFPN Architecture

Input: Feature maps set $C = \{C_3, C_4, C_5, C_6, C_7\}$ from different levels of the backbone CNN
Output: Enhanced feature maps set $P = \{P_3, P_4, P_5, P_6, P_7\}$

```
function GS-BiFPN($C$)
    $P \leftarrow \{\}$
    for $i = 3$ to 7 do
        $P_i \leftarrow Conv_{1\times1}(C_i) + UpSample(P_{i+1})$
    end for
    for $i = 7$ downto 3 do
        if $i > 3$ then
            $P_i \leftarrow GSConv(Merge(P_i, DownSample(P_{i-1})))$
        end if
    end for
    $P_3 \leftarrow UpSample(P_4)$
    $P_5 \leftarrow GSConv(Merge(C_5, DownSample(P_3), P_5, UpSample(P_6)))$
    $P_6 \leftarrow GSConv(Merge(C_6, DownSample(P_3), P_6, UpSample(P_7)))$
    $P_7 \leftarrow GSConv(Merge(C_7, DownSample(P_3), P_7))$
    return $P$
end function
```

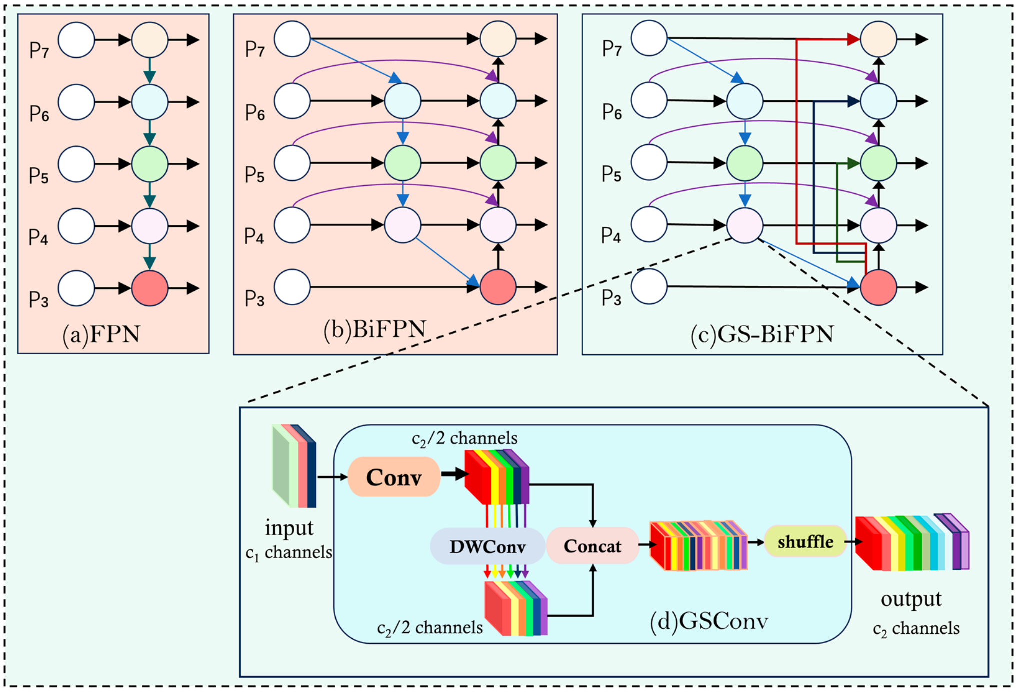

Figure 5. Comparison of feature pyramid network structures. (**a**) Displays the traditional FPN structure. (**b**) Reveals the enhanced strategy of BiFPN, which introduces additional top–down and bottom–up paths to foster richer interaction among features. (**c**) Is our proposed GS-BiFPN, which further enhances feature flow by direct connections and multi-point fusion to strengthen the propagation efficiency of multi-scale features. (**d**) Details the design of GSConv, a component that reduces computational complexity through grouped convolution and channel shuffling techniques while maintaining effective feature fusion capabilities.

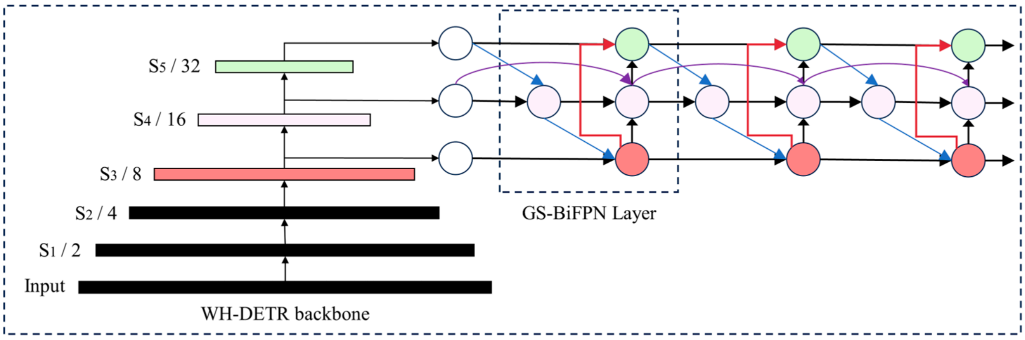

Figure 6. Specific application of GS-BiFPN in the WH-DETR detection architecture.

The operation of standard convolution can be described as performing a fully connected convolution on the input feature map I using the kernel K. The computational cost of this operation can be estimated using the following formula:

$$O_{std} = Conv(I, K) \tag{1}$$

Considering that the dimensions of the input feature map are $H \times W \times C_{in}$, the size of the convolution kernel is $D_k \times D_k$, and the number of output feature map channels is C_{out}, the computational complexity of the standard convolution can be approximately calculated as:

$$Time(O_{std}) = H \times W \times D_k^2 \times C_{in} \times C_{out} \tag{2}$$

Compared to standard convolution, GSConv significantly reduces computational complexity and parameter count by grouping the input feature map and performing convolution independently within each group, followed by channel shuffling. The computational complexity of GSConv can be estimated using the following formula:

$$O_{gs} = Shuffle(GroupConv(I, K, G)) \tag{3}$$

If the input feature map is divided into G groups, with the number of channels per group becoming C_{in}/G, the computational complexity of GSConv can be expressed as:

$$Time(O_{gs}) = \frac{1}{G} \times H \times W \times D_k^2 \times C_{in} \times C_{out} \tag{4}$$

Adding channel shuffling does introduce additional computation; however, since this computation is relatively minor, overall, GSConv can significantly reduce computational complexity—approximately by a factor of G (where G is the number of groups)—while maintaining a similar representational capability compared to standard convolution.

In the Feature Pyramid Network (FPN), the input feature maps C_i come from different levels of the underlying convolutional neural network, representing multi-scale visual features from shallow to deep layers. We have a feature pyramid P_i, where i denotes different pyramid levels. For each specific level i, the input feature map C_i undergoes feature fusion to produce the output feature map P_i. C_i is typically reduced in channel number by a 1 × 1 convolution layer, then merged with the feature map of the next layer through an upsampling step. The formula is expressed as:

$$P_i = Conv_{1\times1}(C_i) + UpSample(P_{i+1}) \tag{5}$$

$Conv_{1\times1}$ is a 1 × 1 convolution used to reduce the number of channels, facilitating the addition of the upsampled feature map P_{i+1}. The $UpSample$ operation denotes upsampling.

The BiFPN enhances the information pathway from top to bottom, and for each pyramid level i, it integrates information not only from the previous layer, but also considers the information from lower-resolution levels.

$$P_i^{'} = Conv\Big(Merge\Big(C_i, UpSample\Big(P_{i+1}^{'}\Big), DownSample\Big(P_{i-1}^{'}\Big)\Big)\Big) \tag{6}$$

where $Merge$ refers to feature fusion, incorporating information from various resolution levels.

In GS-BiFPN, to mitigate the degradation of detailed information as it propagates through deep networks, direct connections from the P3 layer to lower-resolution feature maps P5, P6, and P7 are added to enhance the transfer of detail features. Additionally, GSConv is introduced to reduce the computational costs of feature fusion. The formulas for the fused feature maps P5, P6, and P7 are as follows:

$$P_5^{''} = GSConv\big(Merge\big(C_5, DownSample\big(P_3^{''}\big), UpSample\big(P_6^{''}\big), DownSample\big(P_4^{''}\big)\big)\big) \tag{7}$$

$$P_6^{''} = GSConv\big(Merge\big(C_6, DownSample\big(P_3^{''}\big), UpSample\big(P_7^{''}\big), DownSample\big(P_5^{''}\big)\big)\big) \tag{8}$$

$$P_7^{''} = GSConv\big(Merge\big(C_7, DownSample\big(P_3^{''}\big), DownSample\big(P_6^{''}\big)\big)\big) \tag{9}$$

Here, we incorporate information from the P3 layer into the P5, P6, and P7 layers to better preserve the detailed information of wheat spikes.

2.5. Precision Agriculture Visual Perception: Integrating EIoU Loss Function to Optimize Detection of Highly Overlapping Wheat Spikes

In the task of object detection, the design of the loss function is crucial for the performance of the model. It affects not only the efficiency of the model learning, but also determines the accuracy of the detection. Although the GIoU loss function of our base model, RT-DETR, enhances the geometric consistency in object detection, it still faces challenges in handling overlapping objects. GIoU loss considers the overlap and size difference between the predicted and actual bounding boxes, but it does not sufficiently penalize deviations in the shape and position of the boxes. Particularly in agricultural settings, it is critical to accurately distinguish and locate closely overlapping wheat spikes, and the limitations of GIoU may lead to inaccurate estimations in overlapping areas. To address this, we incorporate the EIoU loss function in the WH-DETR model to remedy these shortcomings. The EIoU loss includes considerations for aspect ratio and centroid deviations, guiding the model more meticulously in distinguishing overlapping objects. By incorporating this loss, the model provides more precise feedback on the geometric properties of bounding boxes, especially in cases of overlapping wheat spikes, thus enhancing the detection accuracy and model convergence speed. The definition of the EIoU loss function is as follows:

$$L_{EIoU} = 1 - IoU + \rho^2\left(b, b_{gt}\right) + \alpha\rho^2\left(w, w_{gt}\right) + \beta\rho^2\left(h, h_{gt}\right) \tag{10}$$

Here, IoU denotes the intersection over union between the predicted and actual bounding boxes. $\rho^2\left(b, b_{gt}\right)$ represents the squared Euclidean distance between the centroids of the predicted and actual boxes, accounting for positional deviations and penalizing inaccuracies in centroid localization. α and β are weighting coefficients used to balance the size loss terms. $\rho^2\left(w, w_{gt}\right)$ and $\rho^2\left(h, h_{gt}\right)$, respectively, represent the squared differences in width w and height h between the predicted and actual boxes. These components specifically address errors in target size and shape to enhance the detection accuracy of small objects and those with significant shape variations.

In the WH-DETR model, we further integrated the EIoU loss to enhance the model's adaptability to complex scenes. The integrated loss function is as follows:

$$L_{WH-DETR} = L_{RT-DETR} + \lambda_{EIoU}L_{EIoU} \tag{11}$$

In the WH-DETR model, $L_{RT-DETR}$ represents the loss function of the RT-DETR network, L_{EIOU} includes additional geometric considerations from EIoU, and λ_{EIoU} is a weighting factor used to balance the impact of the EIoU loss within the overall loss function. By incorporating the EIoU loss, WH-DETR achieves faster convergence and a higher localization accuracy in scenarios that require a nuanced understanding of the geometric relationships between predicted and actual bounding boxes.

3. Results and Discussion

3.1. Experiment Settings

The experiments in this study were conducted using the PyTorch 2.0.0 framework on the following hardware environment: NVIDIA RTX 3090 GPU (24 GB of VRAM) and Intel(R) Xeon(R) Gold 6330 CPU (14 vCPUs, 2.00 GHz) with 80 GB of system memory. The software depends on Python 3.8 and Cuda 11.8.

Each experiment was initiated with weights that were randomly assigned, without reliance on pre-trained models throughout the training process. The dataset, formatted in YOLO style, was partitioned into an 80% training subset and a 20% testing subset, including 5,400 training and 1,350 testing images. The training lasted for 100 epochs, with images maintained at a resolution of 1024 × 1024 pixels and batches composed of eight images. Optimization was managed by the Adam algorithm, starting with a learning rate of 1×10^{-4}, which was progressively adjusted using a cosine annealing approach. These settings were carefully chosen to optimize the training efficacy and ensure consistent experimental outcomes across tests.

3.2. Evaluation Indicator

In our research, we assessed the model's efficacy using metrics including AP, recall, FPS, and model dimensions. AP represents the area under the curve of the precision–recall relationship, indicating the model's average efficacy throughout the dataset. Recall measures the ratio of accurately identified wheat spikes against all actual wheat spikes present, and precision measures the ratio of accurately identified wheat spikes against all spikes detected by the model. Below are the formulas used for these calculations:

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{TP + FN} \tag{13}$$

$$AP = \int_0^1 precision(recall)d(recall) \tag{14}$$

TP (True Positives), FP (False Positives), and TN (False Negatives) represent the number of wheat spikes correctly detected, incorrectly detected, and missed, respectively. In the field of object detection, the AP metric is widely used to evaluate the comprehensive detection performance of models. To quantify the detection speed, we calculated the frames per second (FPS) of the model, with the formula as follows:

$$FPS = \frac{1}{T} \tag{15}$$

where T represents the average time required to process a single frame image.

In this study, we used AP50 as the primary evaluation metric, which measures the proportion of correct positive detections by the model at an IoU threshold of 0.5. Additionally, the real-time performance of the model was assessed by the frames per second (FPS) metric, which reflects the model's capability to process input images in real time.

3.3. Model Performance

In this experimental phase, we thoroughly assessed the WH-DETR model's capability to detect wheat spikes. This comparison aimed to elucidate the performance variations among different network models and specifically underscore WH-DETR's superior accuracy and speed.

In this study, we evaluated the WH-DETR model for wheat spike detection and compared it against nine leading object detection algorithms, as summarized in Table 1. These included two-stage networks such as Faster R-CNN [21] and Mask R-CNN [45], one-stage networks like SSD [27], Centernet [29], YOLOv5 [24], and YOLOv8 [26], and Transformer-based variants like Deformable DETR [33], DINO-Deformable-DETR [46], and our baseline, RT-DETR [35]. WH-DETR achieved a score of 0.957 on the standard AP50 metric, surpassing these existing technologies. Particularly notable in terms of its parameter efficiency, despite YOLOv8 having more parameters (43 M compared to 37 M) and a slightly higher FPS (58.72 compared to 46.68), WH-DETR still outperformed in the AP50 metric, achieving 0.957, surpassing YOLOv8's 0.912. Additionally, compared to Deformable DETR and DINO-Deformable-DETR, with AP50 scores of 0.878 and 0.895, respectively, WH-DETR established a new benchmark for performance. In terms of real-time processing capabilities, despite having fewer parameters (37 M vs. 47 M for DINO-Deformable-DETR), WH-DETR achieved a high FPS of 46.68, significantly enhancing real-time processing capabilities compared to 18.36. This advantage not only demonstrates WH-DETR's meticulous balance of efficiency and performance, but also implies that, in resource-limited deployment environments, WH-DETR can operate more efficiently. Additionally, in Figure 7, the training loss curve comparisons showed significant performance advantages for WH-DETR, with faster and more stable loss reduction, validating the critical role of optimized loss functions in enhancing detection performance. The confusion matrix in Figure 8 further illustrates the performance of WH-DETR in distinguishing between wheat heads and background. The high accuracy of wheat head detection, with minimal misclassification, emphasizes the robustness and reliability of the WH-DETR model.

Table 1. Detection results on GWHD2021. DINO-Deformable-DETR is abbreviated as DINO.

Model	Type	Backbone	AP_{50}	$FPS_{bs=1}$	Params (M)
SSD [27]	One-Stage	VGG-16	0.784	51.73	16
Centernet [29]		ResNet-50	0.832	49.52	29
YOLOv5-L [24]			0.907	**60.32** [1]	46
YOLOv8-L [26]			0.912	58.72	43
Faster R-CNN [21]	Two-Stage	ResNet-50	0.856	12.83	41
Mask R-CNN [45]		ResNet-50	0.862	13.17	44
Deformable DETR [33]	DETR	ResNet-50	0.878	15.27	40
DINO [46]		ResNet-50	0.895	18.36	47
RT-DETR-L [35]		HGNetv2	0.914	34.52	32
WH-DETR (Ours) [2]		HGNetv2	**0.957** [1]	46.68	37

[1] Bold values indicate statistically significant results. 2 indicates our proposed model.

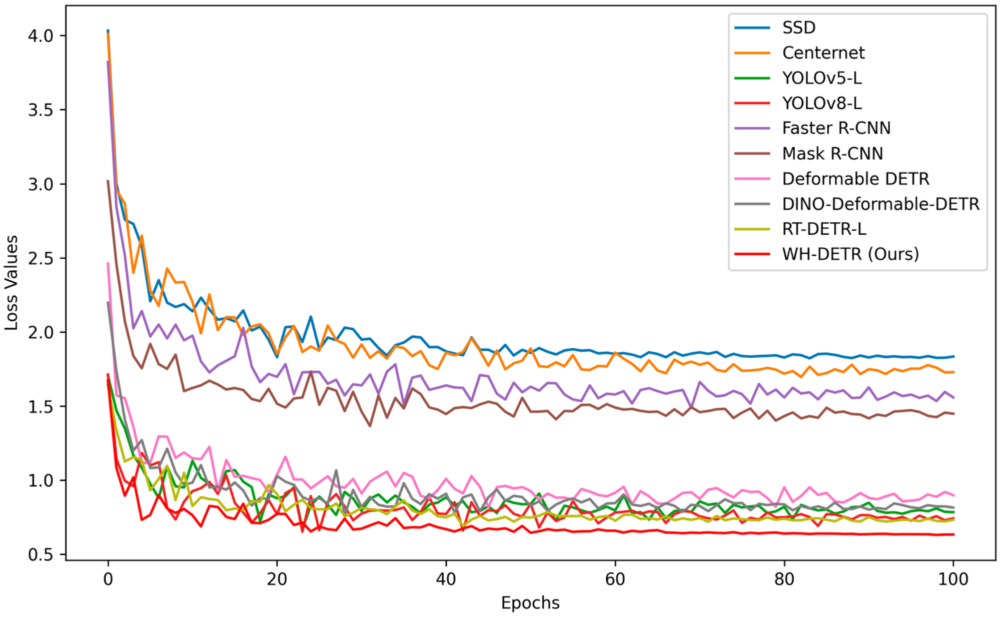

Figure 7. Comparison of loss function trends across different detection models.

In Figure 9, we conducted a comparative analysis of the detection performance of the WH-DETR, YOLOv8, Faster R-CNN, and RT-DETR models. Particularly in challenging lighting conditions, the WH-DETR model maintained a high detection accuracy, especially evident in images C and D. In these images, the model accurately identified wheat spikes under both strong and weak lighting, highlighting WH-DETR's capability to adapt to varying lighting conditions. In image C, under low light, Faster R-CNN, YOLOv8, and RT-DETR all exhibited varying degrees of missed detections, indicating WH-DETR's superior robustness under uneven lighting conditions. In image D, under strong lighting, the misdetections by Faster R-CNN and missed detections by YOLOv8 and RT-DETR further demonstrated WH-DETR's stability in extreme conditions. These comparative results might point to WH-DETR's innovations in feature fusion strategies and loss function design, which likely support the model's stable detection performance under different lighting conditions. In image A, although all models accurately identified most wheat spikes, WH-DETR performed better in detecting spikes at the edges and under slight obstructions, indicating its more precise boundary feature capture. Furthermore, image B showed WH-DETR's higher robustness to variations in the shape and size of wheat spikes, possibly due to its loss function design, enhancing the model's recognition of small-sized and shape-changing targets. Image E showcased the detection of wheat spikes against a very uniform background, where WH-DETR avoided overfitting to background noise, a weakness observed in other models. This robustness against background interference may stem from the diverse data augmentation techniques applied during WH-DETR's training process.

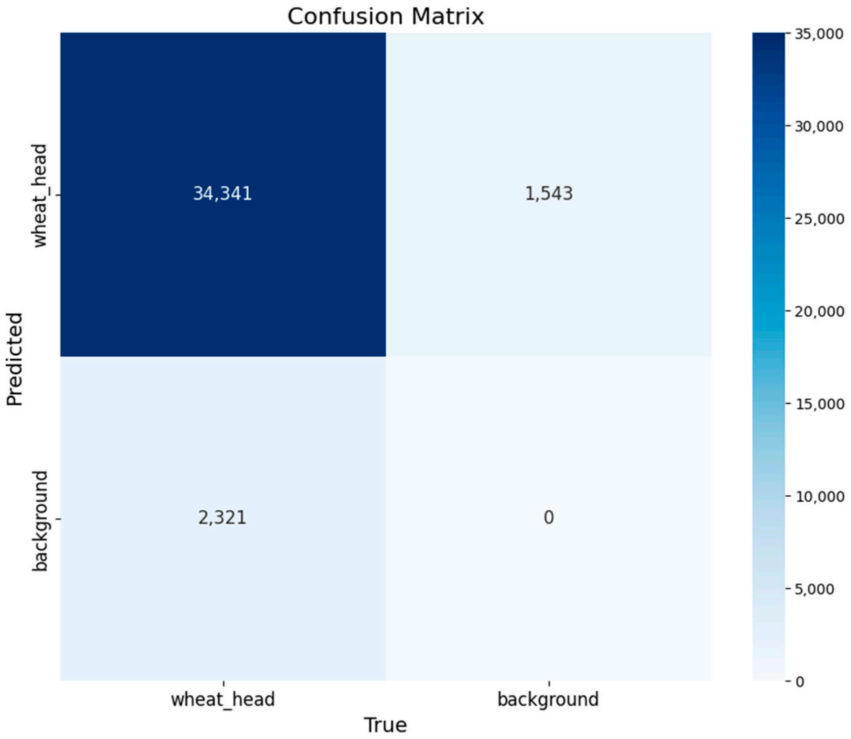

Figure 8. Confusion Matrix of WH-DETR model performance.

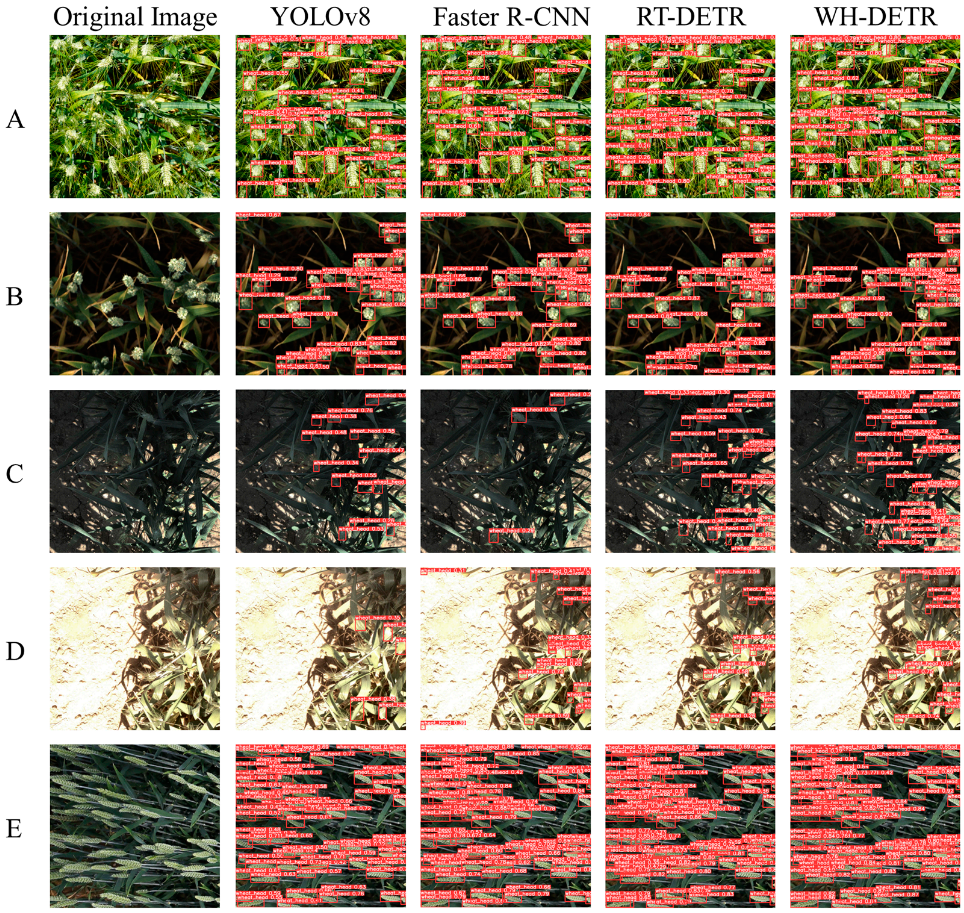

Figure 9. Wheat spike detection algorithm comparison: this set of images demonstrates the performance differences of the algorithm in detecting wheat spikes under various environmental conditions.

(**A**) Depicts the detection scenario under severe background interference; (**B**) shows the detection outcome in sparse vegetation; (**C**) reflects the algorithm's performance in low-light conditions; (**D**) demonstrates the detection effectiveness under strong lighting; and (**E**) displays the detection results in high vegetation density.

In summary, the WH-DETR model exhibited significant performance advantages under varying environmental conditions, especially in scenarios with uneven lighting and complex obstructions. Its advanced feature processing capabilities and sensitivity to details ensured a high accuracy in real field conditions, while also demonstrating an excellent robustness against background interference. These attributes make the WH-DETR model a strong candidate for precision agriculture and automated plant protection applications.

In Figure 10, we can visually examine the focal points of wheat spike detection by the WH-DETR, YOLOv8, Faster R-CNN, and RT-DETR models through a comparison of heatmaps with the original images. Notably, WH-DETR's heatmap displayed a high concentration and precise recognition of wheat spikes, in stark contrast to the larger error hotspots of YOLOv8 and Faster R-CNN, with WH-DETR's key target highlights being more distinct and concentrated—a visual alignment with the actual distribution of wheat spikes in the original images. This concentration indicated our model's precision in feature extraction and accuracy in target localization. While RT-DETR showed more extensive coverage in the detection area, it appeared to be less refined in recognizing specific complex scenes. As shown in Figure 10a,c, RT-DETR struggled to accurately focus on wheat spike features within extensive detection areas, increasing the likelihood of false detections. Conversely, WH-DETR demonstrated precise attention to the areas where wheat spikes were distributed, thus achieving a higher detection accuracy. As illustrated in Figure 10b,d, WH-DETR precisely focused on scenarios where wheat spikes overlapped, enhancing detection efficiency. Although other areas received less attention, the spikes there were in less disturbed regions, making them easier to detect. WH-DETR's heatmap not only reflects its high sensitivity to small targets against complex backgrounds, but also reveals its high specificity in target detection while maintaining a low false detection rate, a critical feature for detecting wheat spikes under dense coverage and obstructive conditions, further highlighting the significant progress made by the WH-DETR model.

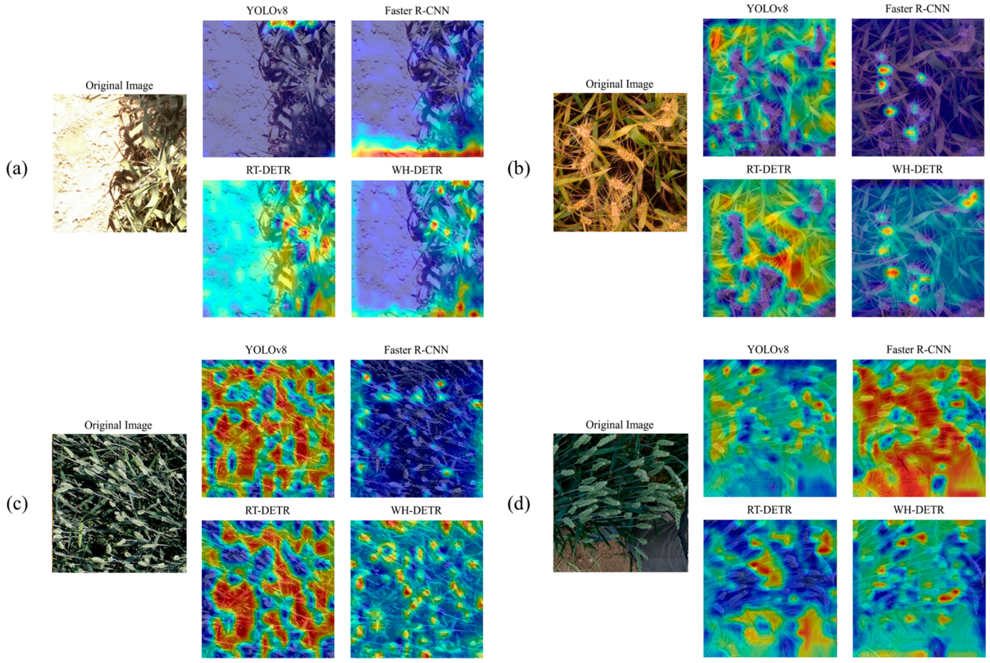

Figure 10. (**a**–**d**) In the figure are examples of heat maps: for each heat map, the top left image is the original wheat spike image, while the top right and bottom row images display the corresponding

heatmaps. These heatmaps represent the response intensity of different algorithms to the target features. In the heatmaps, red areas indicate the algorithm's predicted focus areas, i.e., the possible locations of wheat spikes, while blue areas denote regions of lower focus.

3.4. Ablation Experiments

Before delving into an in-depth analysis of the WH-DETR model, we first highlight our innovations by comparing them with the baseline model, RT-DETR. As our research starting point, RT-DETR has already established a solid foundation in object detection tasks. However, the WH-DETR model, by incorporating three key technologies—GS-BiFPN, GSConv, and EIoU loss function—not only significantly improved the AP50 index from 0.914 to 0.957 in a similar parameter count, but also demonstrated a superior real-time performance (FPS increased to 46.68). This comparison provides a clear research motivation for our subsequent ablation studies: to systematically identify the specific contributions of each innovation to the model's performance and assess the impact of each improvement on the model performance.

3.4.1. The Impact of the Lightweight Feature Pyramid Network GS-BiFPN

First, the introduction of the lightweight feature pyramid network, GS-BiFPN, was evaluated, as detailed in Table 2. To assess the contribution of GS-BiFPN to the model performance, we initially removed the GS-BiFPN module while keeping all other modules unchanged, then sequentially reintroduced FPN, BiFPN, and GS-BiFPN. The results showed that removing GS-BiFPN led to a decrease in AP50 from 0.957 to 0.914, while FPS increased from 46.68 to 50.92. This indicated that GS-BiFPN was crucial for effective integration of multi-scale features, improving AP50 by 4.3%, with only a minor decrease in FPS. Compared to introducing FPN and BiFPN, while FPS was similar among the three, the introduction of GS-BiFPN resulted in 5.4% and 3.5% increases in AP50 over FPN and BiFPN, respectively.

Table 2. Comparison of the specific impacts of GS-BiFPN and other feature pyramid network architectures on Average Precision (AP50), Frame Rate (FPS), and Model Parameter Count in the WH-DETR model.

Model	AP_{50}	$FPS_{bs = 1}$	Params (M)
WH-DETR without GS-BiFPN	0.914	50.92	28
WH-DETR with FPN	0.903	47.34	34
WH-DETR with BiFPN	0.922	45.52	36
WH-DETR with GS-BiFPN	**0.957** [1]	46.68	37

[1] Bold values indicate statistically significant results.

The introduction of GS-BiFPN played a decisive role in improving the WH-DETR model's AP50 performance to 0.957. This significant enhancement can be attributed to the optimization of feature fusion and scale perception capabilities by GS-BiFPN. By enabling high-level feature maps to directly utilize the detailed information flow from lower-level feature maps, the model significantly improved its detection accuracy for targets of various sizes, particularly enhancing its capability to capture fine-grained features. This feature fusion strategy, which incorporates grouped convolution and depth-wise separable convolution, not only optimized the integration process of the information flow, but also enhanced the efficiency and diversity of the feature extraction. Although the introduction of GS-BiFPN increased the computational demands, the real-time processing speed of the model—the FPS slightly decreased from 50.92 to 46.68—still remained at a high level. This demonstrated that GS-BiFPN's design managed to enhance performance while also considering the efficiency of computational resource usage. In real-time application scenarios, this aspect is especially crucial, as it ensures that the model operates efficiently without sacrificing detection accuracy.

Compared with other feature pyramid networks like FPN and BiFPN, GS-BiFPN's performance advantage was further confirmed. Our analysis revealed that GS-BiFPN not only strengthened the model's feature representation capabilities, but also balanced the relationship between performance and computational efficiency. Additionally, results from ablation studies highlighted the critical role of GS-BiFPN in the design of WH-DETR and provided valuable insights for future efforts to find a better balance between performance and efficiency in model optimization.

3.4.2. The Contribution of Introducing GSConv

Next, we separately evaluated the improvements brought about by the introduction of GSConv, as detailed in Table 3. Initially, we removed GSConv and then reintroduced it in the overall network, the head part, and the neck part. When the GSConv module was removed, we observed a slight increase in AP50 performance; however, the FPS significantly dropped to 33.53. When GSConv was globally reintroduced, there was a slight improvement in FPS (48.14), but AP50 decreased by 3.6% compared to when GSConv was introduced only in the neck part. When introduced solely in the head part, the overall performance declined.

Table 3. Comparison of the impact on metrics from different integration strategies of GSConv in the WH-DETR model.

Model	AP_{50}	$FPS_{bs = 1}$	Params (M)
WH-DETR without GS-Conv	**0.959** [1]	33.53	45
WH-DETR with GSConv(ALL)	0.921	**48.14** [1]	33
WH-DETR with GSConv(HEAD)	0.903	45.71	36
WH-DETR with GSConv(NECK)	**0.957** [1]	46.68	37

[1] Bold values indicate statistically significant results.

The analysis showed that, without GSConv, the model reached an AP50 of 0.959—a relatively high benchmark, but at the cost of a lower FPS (33.53). This reflects that traditional convolution operations, while performing well, exhibit noticeable limitations in processing speed, especially in scenarios requiring real-time responses. On the other hand, when GSConv was incorporated throughout the model, despite a significant reduction in the parameter count to 33 M, the FPS improved markedly to 48.14, but AP50 decreased to 0.921. This indicated that, while GSConv was effective in reducing the model complexity and increasing the processing speed, its grouped and depth-wise separable features may have reduced the granularity of the feature extraction, thus impacting the overall detection accuracy. This global application might remove key information useful for subsequent tasks in the early layers of feature extraction or overly simplify the processing steps where a higher computational power is needed. When GSConv was introduced in the neck part of the model, it maintained a high accuracy (AP50 of 0.957) with a reasonable speed (46.68 FPS), indicating that the neck part was a critical area for GSConv's effectiveness. Specifically, the neck part was responsible for transforming high-level abstract features extracted from the backbone into richer and more diverse feature representations, which are crucial for subsequent object detection tasks. GSConv, through its grouping and depth-wise separable mechanisms, maintained a sufficient feature representation capacity while improving computational efficiency. This might explain why the introduction of GSConv in the neck part of WH-DETR could improve the AP50, and since the neck part acts as a "bridge" in feature transmission, optimizing this part had a significant impact on the overall model performance. Conversely, when GSConv was applied to the head part of the model, the accuracy decreased (AP50 of 0.903), indicating that the head part required more complex feature-processing capabilities, and the simplified operations of GSConv might limit performance in this section. The head part typically involves the direct decoding and classification of multiple feature layers, requiring a higher level of fine-grained feature expression. Therefore, GSConv in this part may not provide sufficient feature retention, leading to a slight reduction in accuracy. Considering the number of model parameters, we see that the use of GSConv significantly reduced the parameter count, which is particularly valuable in resource-constrained application scenarios.

Overall, when GSConv was applied only in the neck part, the model almost matched the precision of the version without GSConv (AP50 of 0.957), while improving speed (FPS of 46.68), and the parameter count increased only slightly to 37 M. This indicated that applying GSConv in this specific part of the model found an appropriate balance between efficiency and effectiveness. The neck part, as a key module linking the backbone and head, was crucial for enhancing the overall model performance without sacrificing the necessary feature details.

3.4.3. The Effects of Integrating the EIoU Loss Function

Finally, this study explores the impact of integrating the EIoU loss function on the performance of the WH-DETR model, as detailed in Table 4. The experiments included scenarios without incorporating any other loss functions, as well as introducing CIoU, WIoU, and EIoU, respectively.

Table 4. Results of WH-DETR with different IoU loss functions.

Method	**AP$_{50}$**	**Recall**
Original Loss(GIoU)	0.922	0.883
Original Loss + CIoU	0.925	0.892
Original Loss + WIoU	0.923	0.895
Original Loss + EIoU [1]	**0.957** [1]	0.927

[1] Indicates the proposed model (WH-DETR) and the use of the Original Loss + EIoU configuration.

The experimental results indicate that, among all the tested loss function variants, the addition of the EIoU loss function (Original Loss + EIoU) yielded the most significant performance enhancement to the model, with AP50 increasing from 0.922 to 0.957 and Recall from 0.883 to 0.927. This outcome strongly suggests the critical role of the EIoU loss function in precise target localization during model training, particularly in enhancing the model's sensitivity to IoU overlap scenarios. The superiority of the EIoU loss function may result from its meticulous consideration of target size and shape, which allows for more accurate guidance on how the model learns to align with the actual targets. Compared to the GIoU loss (Original Loss + GIoU) and other variants such as CIoU and WIoU, EIoU provides a more complex error signal, aiding in more accurate predictions of the target bounding boxes in scenarios with obstructions and overlaps. Furthermore, analyses showed that, while CIoU (Original Loss + CIoU) and WIoU (Original Loss + WIoU) also achieved performance improvements in AP50 and Recall, the increases were smaller. This may indicate the limitations of these loss functions in specific object detection scenarios or their relatively incomplete consideration of target scale, shape, and contextual information compared to EIoU. The importance of loss functions in training deep learning models is crucial, as they directly affect how models adjust weights during the optimization process to reduce prediction errors. The effectiveness of the EIoU loss function emphasizes the importance of loss function design in object detection tasks and shows that refined loss functions can significantly enhance model accuracy and robustness.

Figure 11 demonstrates the impact of different loss function combinations on wheat spike detection in scenarios with significant overlap. In Figure 11a, we selected a scene with notable overlap (indicated by the red box) for analysis. Comparing Figure 11b–e, the GIoU + EIOU combination excelled in addressing the challenge of overlap, effectively recognizing and reducing redundant detection boxes. This combination particularly stood out in the Class Activation Maps (CAMs) shown in Figure 11f–i, where it concentrated significant attention on the correct locations of wheat spikes, showcasing superior recognition capabilities over other combinations.

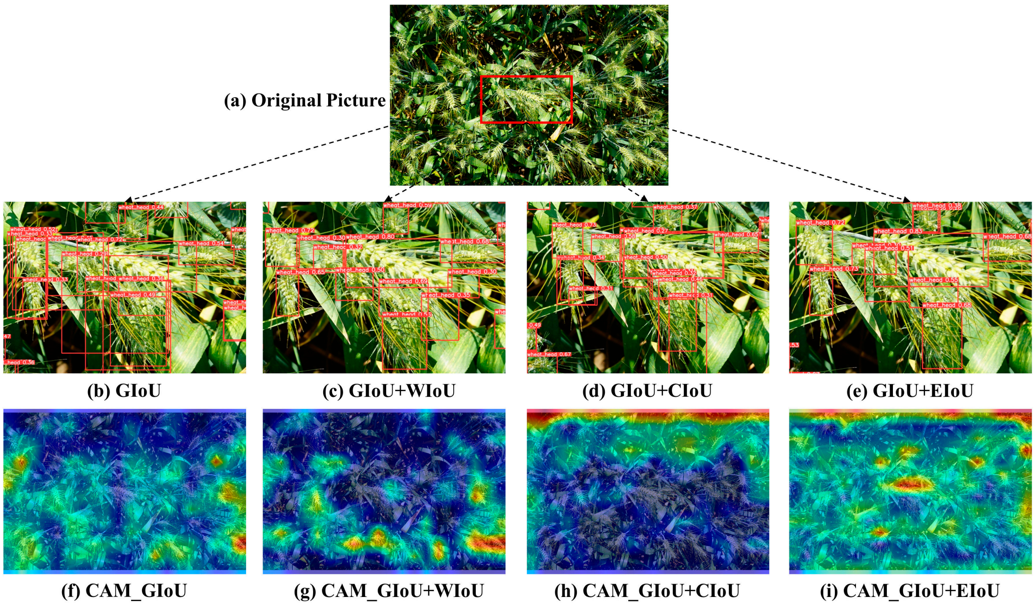

Figure 11. WH-DETR model outputs using different loss functions. (**a**) Original image. (**b**–**e**) Detection results with GIoU, GIoU + WIoU, GIoU + CIoU, and GIoU + EIou losses, respectively. (**f**–**i**) Corresponding CAMs for each loss function.

Moreover, the GIoU and GIoU + CIoU combinations produced a higher number of overlapping detection boxes in these complex scenes, indicating a substantial detection redundancy. Although the GIoU + WIoU combination showed some improvement, it still fell short in accuracy and the control of redundancies. The superior performance of the GIoU + EIOU combination can be attributed to the introduction of the EIOU loss, which effectively optimized the alignment of bounding boxes and the interaction between targets, especially in overlapping scenarios. The EIOU loss enhanced the detection model's sensitivity to overlapping areas of the bounding boxes, thereby improving the detection precision and significantly reducing redundancies. In contrast, although other loss functions considered the geometric alignment of targets, their optimization effects were less pronounced in complex scenarios with multiple overlapping targets. These observations provide crucial experimental evidence for selecting loss functions in wheat spike detection models, particularly in agricultural environments with frequent target overlaps.

Overall, the integration of the EIoU loss function significantly improved the WH-DETR model's performance on the AP50 and Recall metrics, highlighting the importance of considering detailed target localization errors in loss function design.

3.4.4. Summary of Ablation Studies

The results of our ablation studies indicated that both GSConv and GS-BiFPN played crucial roles in enhancing the performance of the WH-DETR model. GSConv significantly reduced the computational complexity, which contributed to the increased FPS observed in our experiments. Meanwhile, the introduction of GS-BiFPN improved the efficiency of feature fusion, leading to a better detection accuracy. Although the inclusion of GS-BiFPN increased the parameter count, its impact on real-time processing was mitigated by the efficiency gains from GSConv. This combined effect explains why the WH-DETR model, despite having more parameters than RT-DETR_L, achieved a higher FPS. Additionally, the integration of the EIoU loss function significantly improved the AP50 and Recall metrics, emphasizing the importance of detailed target localization. These findings underscore the importance of balancing model complexity and computational efficiency to optimize performance.

4. Conclusions

This study introduced an improved RT-DETR model, WH-DETR, aimed at enhancing performance in wheat spike detection tasks. Through in-depth experimental analysis, we demonstrated the significance of integrating GS-BiFPN, GSConv, and the EIoU loss function in the model for enhancing the detection accuracy in complex scenarios. The WH-DETR model achieved outstanding results across various metrics, and the improvement in AP50 validates our strategies for feature fusion and loss function refinement. The innovations of this model are not only in performance enhancement, but also in optimizing real-time processing capabilities and computational efficiency. This provides significant convenience for practical applications needing rapid responses in resource-limited field environments, such as drone monitoring and automated harvesting. Additionally, the scalability of WH-DETR lays the groundwork for future applications in other crop detection tasks. In summary, the WH-DETR model represents a significant step forward in advancing target detection technology in precision agriculture, offering substantial technical support for achieving a more efficient and intelligent agricultural production system. Despite significant progress in wheat spike detection with WH-DETR, we identify the following directions as crucial for future research:

- Model Light-weighting and Acceleration: Further research into model light-weighting and acceleration techniques to adapt to edge computing devices, promoting the model's application in actual agricultural production environments.
- Multitask Learning: Explore multitask learning models that integrate wheat spike detection with other agricultural tasks, such as pest and disease identification and growth stage prediction, to achieve more comprehensive agricultural monitoring.
- Extending Application Scope: Investigate the applicability of the WH-DETR model in a broader range of target detection tasks, including adaptability to different environmental conditions and crop types, to enhance the diversity and robustness of agricultural monitoring systems.

Author Contributions: Z.Y.: conceptualization, methodology, software, investigation, formal analysis, and writing—original draft, data curation, project administration and validation. W.Y.: methodology, validation, formal analysis, visualization and writing—original draft. J.Y.: conceptualization, funding acquisition, resources, validation, supervision, and writing—review and editing. R.L.: funding acquisition, supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the General Program Hunan Provincial Natural Science Foundation of 2022, China (2022JJ31022). The funder is the Department of science and technology of Hunan Province.

Institutional Review Board Statement: The study is focused on wheat spike detection and does not involve humans or animals. Therefore, ethical review and approval are not applicable to this study. Please note that the Institutional Review Board Statement and approval number are not required.

Data Availability Statement: The datasets analyzed for this study can be found in the Global Wheat Head Detection 2021 dataset at http://www.global-wheat.com (accessed on 9 August 2023). The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding authors.

Acknowledgments: The research benefited from support and assistance at various stages. The authors would like to express special thanks to Yuqiu Zhao at the College of Computer Science and Mathematics at Central South University of Forestry and Technology for his invaluable guidance and suggestions throughout the project.

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

References

1. Shewry, P.R. Wheat. *J. Exp. Bot.* **2009**, *60*, 1537–1553. [CrossRef] [PubMed]
2. Igrejas, G.; Ikeda, T.M.; Guzmán, C. *Wheat Quality for Improving Processing and Human Health*; Springer International Publishing: Cham, Switzerland, 2020. [CrossRef]
3. Food and Agriculture Organization of the United Nations. FAO Cereal Supply and Demand Brief | Food and Agriculture Organization of the United Nations. WorldFoodSituation. Available online: https://www.fao.org/worldfoodsituation/csdb (accessed on 14 June 2024).
4. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L.; et al. Integrating Satellite and Climate Data to Predict Wheat Yield in Australia Using Machine Learning Approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159. [CrossRef]
5. Etienne, D. *Global Wheat Head Dataset 2021*; Zenodo: Geneva, Switzerland, 2021. [CrossRef]
6. Madec, S.; Jin, X.; Lu, H.; De Solan, B.; Liu, S.; Duyme, F.; Heritier, E.; Baret, F. Ear Density Estimation from High Resolution RGB Imagery Using Deep Learning Technique. *Agric. For. Meteorol.* **2019**, *264*, 225–234. [CrossRef]
7. Pantazi, X.E.; Moshou, D.; Alexandridis, T.; Whetton, R.L.; Mouazen, A.M. Wheat Yield Prediction Using Machine Learning and Advanced Sensing Techniques. *Comput. Electron. Agric.* **2016**, *121*, 57–65. [CrossRef]
8. Franch, B.; Vermote, E.F.; Skakun, S.; Roger, J.C.; Becker-Reshef, I.; Murphy, E.; Justice, C. Remote Sensing Based Yield Monitoring: Application to Winter Wheat in United States and Ukraine. *Int. J. Appl. Earth Obs. Geoinf.* **2019**, *76*, 112–127. [CrossRef]
9. Rocha, H.; Dias, J. Early Prediction of Durum Wheat Yield in Spain Using Radial Basis Functions Interpolation Models Based on Agroclimatic Data. *Comput. Electron. Agric.* **2019**, *157*, 427–435. [CrossRef]
10. Balasubramanian, V.N.; Guo, W.; Chandra, A.L.; Desai, S.V. Computer Vision with Deep Learning for Plant Phenotyping in Agriculture: A Survey. *Adv. Comput. Commun.* **2020**, arXiv:2006.11391. [CrossRef]
11. Liu, C.; Wang, K.; Lu, H.; Cao, Z. Dynamic Color Transform Networks for Wheat Head Detection. *Plant Phenomics* **2022**, *2022*, 9818452. [CrossRef] [PubMed]
12. Misra, T.; Arora, A.; Marwaha, S.; Chinnusamy, V.; Rao, A.R.; Jain, R.; Sahoo, R.N.; Ray, M.; Kumar, S.; Raju, D.; et al. SpikeSegNet-a Deep Learning Approach Utilizing Encoder-Decoder Network with Hourglass for Spike Segmentation and Counting in Wheat Plant from Visual Imaging. *Plant Methods* **2020**, *16*, 40. [CrossRef]
13. Chandra, A.L.; Desai, S.V.; Balasubramanian, V.N.; Ninomiya, S.; Guo, W. Active Learning with Point Supervision for Cost-Effective Panicle Detection in Cereal Crops. *Plant Methods* **2020**, *16*, 34. [CrossRef]
14. Hasan, M.M.; Chopin, J.P.; Laga, H.; Miklavcic, S.J. Detection and Analysis of Wheat Spikes Using Convolutional Neural Networks. *Plant Methods* **2018**, *14*, 100. [CrossRef] [PubMed]
15. Gong, B.; Ergu, D.; Cai, Y.; Ma, B. Real-Time Detection for Wheat Head Applying Deep Neural Network. *Sensors* **2020**, *21*, 191. [CrossRef] [PubMed]
16. Sun, J.; Yang, K.; Chen, C.; Shen, J.; Yang, Y.; Wu, X.; Norton, T. Wheat Head Counting in the Wild by an Augmented Feature Pyramid Networks-Based Convolutional Neural Network. *Comput. Electron. Agric.* **2022**, *193*, 106705. [CrossRef]
17. Ye, J.; Yu, Z.; Wang, Y.; Lu, D.; Zhou, H. WheatLFANet: In-Field Detection and Counting of Wheat Heads with High-Real-Time Global Regression Network. *Plant Methods* **2023**, *19*, 103. [CrossRef] [PubMed]
18. Yan, J.; Zhao, J.; Cai, Y.; Wang, S.; Qiu, X.; Yao, X.; Tian, Y.; Zhu, Y.; Cao, W.; Zhang, X. Improving Multi-Scale Detection Layers in the Deep Learning Network for Wheat Spike Detection Based on Interpretive Analysis. *Plant Methods* **2023**, *19*, 46. [CrossRef] [PubMed]
19. Zhao, J.; Cai, Y.; Wang, S.; Yan, J.; Qiu, X.; Yao, X.; Tian, Y.; Zhu, Y.; Cao, W.; Zhang, X. Small and Oriented Wheat Spike Detection at the Filling and Maturity Stages Based on WheatNet. *Plant Phenomics* **2023**, *5*, 0109. [CrossRef] [PubMed]
20. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *arXiv* **2013**, arXiv:1311.2524. [CrossRef]
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640. [CrossRef]
23. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
24. Jocher, G.; Chaurasia, A.; Stoken, A.; Borovec, J.; Kwon, Y.; Michael, K.; Fang, J.; Yifu, Z.; Wong, C.; Montes, D.; et al. *Ultralytics/Yolov5: V7.0–YOLOv5 SOTA Realtime Instance Segmentation*; Zenodo: Geneva, Switzerland, 2022. [CrossRef]
25. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-The-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696. [CrossRef]
26. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv8 by Ultralytics. GitHub. Available online: https://github.com/ultralytics/ultralytics (accessed on 9 August 2023).
27. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. *Comput. Vis.ECCV 2016* **2016**, *9905*, 21–37. [CrossRef] [PubMed]
28. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2018**, arXiv:1708.02002. [CrossRef]

29. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. *arXiv* **2019**, arXiv:1904.08189. [CrossRef]
30. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. *arXiv* **2020**, arXiv:1911.09070. [CrossRef]
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
32. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-To-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872. [CrossRef]
33. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-To-End Object Detection. *arXiv* **2021**, arXiv:2010.04159. [CrossRef]
34. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional DETR for Fast Training Convergence. *arXiv* **2021**, arXiv:2108.06152. [CrossRef]
35. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs Beat YOLOs on Real-Time Object Detection. *arXiv* **2023**, arXiv:2304.08069. [CrossRef]
36. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2017**, arXiv:1612.03144. [CrossRef]
37. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-Neck by GSConv: A Better Design Paradigm of Detector Architectures for Autonomous Vehicles. *J. Real-Time Image Process.* **2024**, *21*, 62. [CrossRef]
38. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520. [CrossRef]
39. Zhang, Y.-F.; Ren, W.; Zhang, Z.; Jia, Z.; Wang, L.; Tan, T. Focal and Efficient IOU Loss for Accurate Bounding Box Regression. *arXiv* **2022**, arXiv:2101.08158. [CrossRef]
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030. [CrossRef]
41. Zhou, Q.; Huang, Z.; Zheng, S.; Jiao, L.; Wang, L.; Wang, R. A Wheat Spike Detection Method Based on Transformer. *Front. Plant Sci.* **2022**, *13*, 1023924. [CrossRef] [PubMed]
42. Zhu, J.; Yang, G.; Feng, X.; Li, X.; Fang, H.; Zhang, J.; Bai, X.; Tao, M.; He, Y. Detecting Wheat Heads from UAV Low-Altitude Remote Sensing Images Using Deep Learning Based on Transformer. *Remote Sens.* **2022**, *14*, 5141. [CrossRef]
43. David, E.; Serouart, M.; Smith, D.; Madec, S.; Velumani, K.; Liu, S.; Wang, X.; Espinosa, F.P.; Shafiee, S.; Tahir, I.S.A.; et al. Global Wheat Head Dataset 2021: More Diversity to Improve the Benchmarking of Wheat Head Localization Methods. *arXiv* **2021**, arXiv:2105.07660. [CrossRef]
44. David, E.; Madec, S.; Sadeghi-Tehran, P.; Aasen, H.; Zheng, B.; Liu, S.; Kirchgessner, N.; Ishikawa, G.; Nagasawa, K.; Badhon, M.A.; et al. 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. *arXiv* **2020**, arXiv:2005.02162. [CrossRef]
45. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870. [CrossRef]
46. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.-Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-To-End Object Detection. *arXiv* **2022**, arXiv:2203.03605. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.