WheatNet-Lite: A Novel Light Weight Network for Wheat Head Detection

Sandesh Bhagat[1]*, Manesh Kokare[1], Vineet Haswani[1], Praful Hambarde[2], Ravi Kamble[1]
[1]SGGS Institute of Engineering and Technology, Nanded, India-431606
[2]Indian Institute of Technology, Ropar, India-140001
{2018pec901, mbkokare, 2017bcs602, kambleravi}@sggs.ac.in, 2018eez0001@iitrpr.ac.in

Abstract

Recently, the potential for wheat head detection has been significantly enhanced using deep learning techniques. However, the significant challenges are variation in growth stages of wheat heads, canopy, genotype, and wheat head orientation. Furthermore, the wheat head detection task gets even more complex due to the overlapping density of wheat heads and the blur image due to the wind. For real-time wheat head detection, designing lightweight deep learning models for edge devices is also challenging. This paper proposes a lightweight WheatNet-Lite architecture to enhance the efficiency and accuracy of wheat head detection. The proposed method utilizes Mixed Depthwise Conv (MDWConv) with an inverted residual bottleneck in the backbone. Additionally, the Modified Spatial Pyramidal Polling (MSPP) effectively extracts the multi-scale features. The final wheat head bounding box prediction is achieved using WheatNet-lite Neck by utilizing Depthwise Convolution (DWConv) with a Feature Pyramid structure. It reduces 54.2 M network parameters in comparison to YOLOV3. The proposed approach outperforms the existing state-of-the-art methods with mean average precision (mAP) of 91.32 mAP@0.5 and 86.10 mAP@0.5 on GWHD and SPIKE datasets, respectively, with only 8.2 M parameters. Also, the new ACID dataset is proposed with bounding box annotation with 76.32 mAP@0.5. The experimental results are demonstrated on three different datasets viz. Global Wheat Head Detection (GWHD) , SPIKE dataset, and Annotated Crop Image Dataset (ACID) showing a significant improvement in the wheat head detection with speed and accuracy.

1. Introduction

Wheat has high demand across the globe with a yearly crop production of 700 million tonnes [12]. With the growing population, the demand is most likely to increase [6].

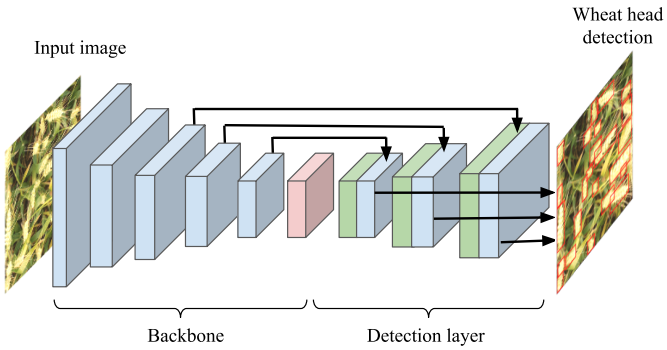

Figure 1. The overall architecture for wheat head detection with modified MobileNetV3 as the backbone and the proposed WheatNet-Lite Neck as detection layer.

The continued development of global wheat is crucial for long-term food security. Wheat production needs to increase by a significant amount to tackle this problem. New plant-breeding techniques allow the development of new wheat plant varieties with desired traits such as disease-resistant, climate-resistant, and higher yields [18]. However, wheat breeding is mainly done in traditional ways, which are almost manual and hence error-prone. The process is tedious and time-consuming. Plant phenotyping measures various properties of plants which may be structural or functional [22]. These techniques are the key to select important wheat traits linked to yield potential, disease resistance, or adaptation to abiotic stress [4]. Among all traits, wheat head numbers per unit ground area is a significant yield component and are still manually evaluated in the breeding trials [4]. Plant breeders use wheat head count in their decision-making process to determine which wheat varieties should be crossed to generate a new, superior offspring.

There is a need for an automated wheat head detection method that helps to mitigate the bottleneck of wheat head detection in wheat breeding. There is a need for an automated wheat head detection method that helps to mitigate the bottleneck of wheat head detection in wheat breeding.

*Corresponding author

(a) (b) (c) (d)

Figure 2. Challenges for wheat head detection (a) Overlapping, (b) Different growth stages, (c) Image blur due to wind (d) Brightness variation

Several studies use computer vision and image processing based methods for wheat head detection[17], [10], [11]. However, there are two types of challenges environmental and wheat inherent challenges. Environmental challenges include blurring due to wind or motion, divergence in the observational conditions, image scale, unwanted shadow and brightening conditions [7]. Wheat inherent challenges such as differences in shapes and colour of wheat head due to genotype variation, different growth stages, overlaps between wheat head, and variation in wheat head orientation make this task more complex. Providing quantitative plant breeder support for yield estimation under actual field conditions, relying on accurately and automatically detecting and counting the wheat head in the field [12]. Due to the wide range of cultivating techniques and appearance, creating an accurate dataset is also a significant challenge. For real-time wheat head detection, designing lite weight deep learning models for edge device applications is also challenging to achieve accurate and efficient wheat head detection. The major contributions of this work are summarized as follows:

1. The proposed method introduces a deep learning technique for wheat head detection using a novel WheatNet-lite network. The extracted features from the proposed backbone show effective representations of the dense and overlapping wheat heads with high accuracy.

2. Automatic extraction of features through Mixed Depthwise Convolution (MDWConv) with an inverted residual bottleneck in the backbone is utilized to enable flexible feature fusion and reduce the parameter in the WheatNet model. Additionally, the modified Spatial Pyramidal Polling (MSPP) effectively extracts the multi-scale features.

3. The wheat head bounding box prediction is achieved using WheatNet-lite Neck by utilizing Depthwise Convolution (DWConv) with a feature pyramid structure drastically reduce parameter.

4. We have proposed an ACID dataset with bounding box annotations. Also, extensive experimentation has been carried out to evaluate the performance of the proposed WheatNet-Lite with state-of-the-art object detection methods such as Fast-RCNN, YOLOV3, and YOLOV4 for the detection of the wheat heads.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 briefly introduces the materials and proposed method for wheat head detection. Section 4 describes the training and implementation, Section 5 gives results and discussion with the Ablation study. Finally, in Section 6, we conclude the paper with ideas for future directions.

2. Related Work

2.1. Convolution Neural Network for wheat head detection

In [21], [7] the image processing techniques is used for wheat head detection. In contrast to traditional image processing techniques, deep learning techniques achieve greater accuracy for wheat head detection.[7], [4], [22], [28], [33], [20], [28]. Convolution Neural Network (CNN) has boosted the performance in various Computer Vision fields such as image classification, image segmentation and object detection. But CNN requires a large amount of training data for and computations to get accurate results. To address this problem, researchers focused on data creation. In [4], a large and diverse GWHD dataset is developed with manual wheat heads annotations, which use as a benchmark dataset for wheat head detection. In [12] , a SPIKE dataset has been created using a camera set up in an oblique view manner which captures a significant number of spike features such as texture, colour, shape, etc. In [22] author focuses on a new ACID dataset creation in the field of wheat phenotyping. For each wheat head in the wheat ears, point annotations are given. In addition, an automated procedure for counting wheat heads has been developed using Light Detection And Ranging's (LiDAR) device in the agricultural field [33]. In [8] the author uses characterization motion-based method in field-grown wheat detection. Two stream CNN architectures for high-resolution image

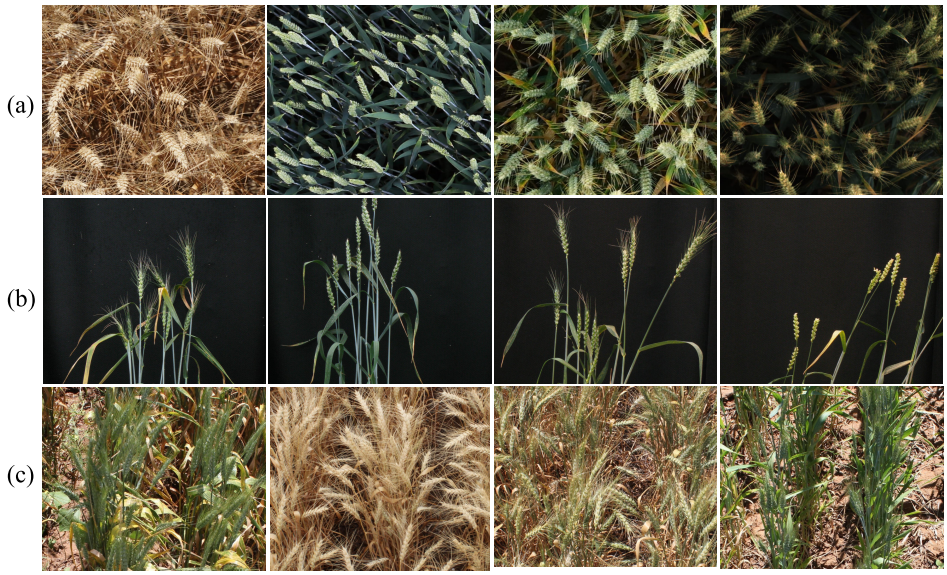

Figure 3. An example of images from datasets: (a) GWHD-Train, (b) ACID and (c) SPIKE dataset.

Table 1. Different datasets used in this study

Dataset	Images collection	Resolution	Number of images	Instances	Annotations type
GWHD-Train [4]	Real field images	1080 x 1080	3422	1,88,445	Bounding box
ACID [22]	Indoor images	1024 x 1024	520	4,100	Bounding box
SPIKE [12]	Real field images	6000 x 4000	335	25,000	Bounding box
Total			8500	45740	

object detection are compared. The first is RCNN, which is faster, and the second is TasselNet, which counts by regression [20]. In deep count [28] paper, linear iterative clustering and deep CNNs based approaches are used to identify and count the number of wheat spikes.

2.2. CNN architecture for Object Detection

CNN based object detection architecture evolved with time with better availability of resources. In [31] author address the problem of multi-scale feature extraction using multiple filter in convolution. In [3] depth-wise separable convolution is proposed to minimize computational complexity while retaining normal convolution performance .

Girshick *et al.* [9] introduced Region-based Convolutional Neural Networks (R-CNN) for object detection using a selective search to detect regions of interest (ROI) and CNN to classify them. Ren *et al.* [27] developed a Faster RCNN employing a region proposal network (RPN) and a CNN for object detection. In [30], efficiency is increased with fewer parameters by scaling the CNN architecture. In [15], the lightweight architecture with inverted bottleneck structure is discussed for efficient feature extraction. Tan *et al.* [32] proposed EfficientDet a combination of bi-directional feature pyramid network and EfficientNet backbones for scalable and efficient object detection. In [23] author proposed You Only Look Once (YOLO) popular CNN based network for object detection. Over time, the YOLO network family has evolved. In [24] work, the YOLOV2 network evolved from the YOLO detection network. Further, in [25] the author proposed the YOLOV3 modified version of YOLOV2. Compared to the Faster R-CNN [26] network, the YOLO network transforms the detection problem into a regression problem. The in-detail literature survey defines a need to increase the efficiency of memory usage and wheat head detection speed to address the demands of real-time applications.

3. Materials and Method

Although many approaches have been proposed in the past, their performance is often not optimal. This paper investigates the uses of recent advanced, efficient deep learning models for accurate wheat head detection.

3.1. Dataset

We evaluate the results of the proposed architecture on three datasets : (1). Global Wheat Head Detection (GWHD) dataset [4], (2). Annotated Crop Image Dataset (ACID) [22], and (3). SPIKE dataset [12], respectively. Sample images from each dataset is shown in Figure 3. Further each of these datasets has been explained in detail as follows:

GWHD dataset [4] is the benchmark for wheat head de-

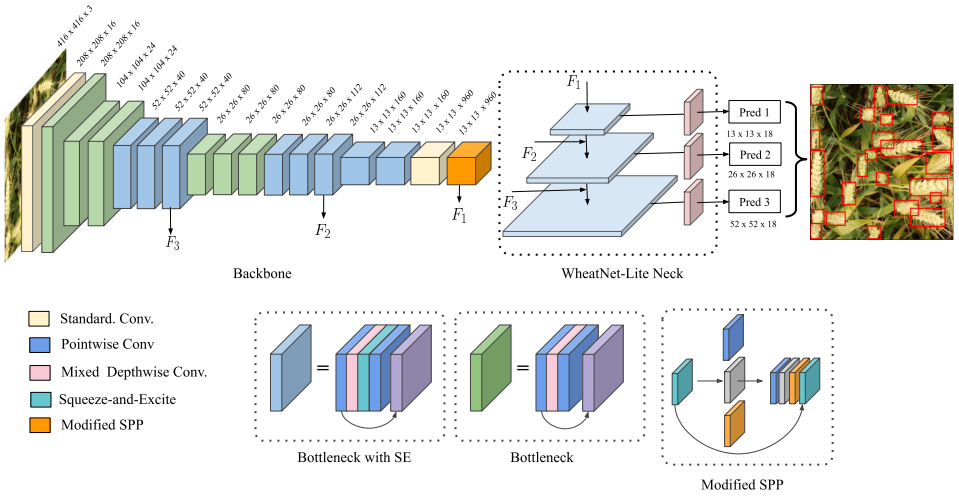

Figure 4. The proposed WheatNet-Lite Architecture.

tection. Data has been collected over four years from 2016 to 2019 from nine research institutions from different locations, with 507 genotypes from Europe, North America, Australia, and Asia [4]. The GWHD dataset includes RGB images collected from a wide variety of field-based phenotyping platforms with cameras. In addition, we collected the GWHD dataset from the global wheat detection competition on the Kaggle platform. As the ground truth of the test dataset is not publicly available, we split the training dataset of 3422 images into training, validation and test dataset with the size of 60% and 20%, and 20%, respectively.

The ACID dataset [22] contains images of wheat plants taken in a glasshouse condition [22]. In total, 520 images are containing a total of 4,100 wheat heads and 48,000 spikelets. Pointwise labels are given for each wheat head, but since we require a rectangular bounding box as annotation for the wheat head detection, which provides more insight than pointwise annotations. Therefore, we manually labeled each wheat head using LabelIMG Annotation tool [2] to create bounding box annotations. Labelled dataset made publically available for research purposes.

The SPIKE dataset [12] has over 300 images of ten wheat varieties at three different growth stages. Annotations for each image presents wheat head bounding boxes. First, the images are directly taken from the field and then cropped, keeping only the Region of Interest (RoI) [12]. Then the images are manually annotated with bounding boxes highlighting all the wheat heads present in the images. The dataset has three such categories: Green Spike and Green Canopy (GSGC), Green Spike and Yellow Canopy (GSYC), Yellow Spike and Yellow Canopy (YSYC). Multiple experts have labelled the images at the resolution of 2000 x 1500 pixels. Each image contains approximately 70–80 spikes. Therefore, in total, the 335 images include about 25,000 annotated wheat heads.

Data pre-processing is a key to achieve high detection accuracy. However, throughout the data analysis, we observe that some wrong bounding boxes are present in the dataset. Thus it can mislead the model while training. To overcome this problem, we remove these wrong bounding boxes from the dataset. The clean dataset is used for training purpose shows significant improvement in the detection accuracy. Different growth stages, image blur due to wind, and variation in the canopy add diversity to a dataset and increase wheat head detection complexity. The proposed method must be robust to these variations of the wheat head. Hence, we have adopted augmentation techniques such as mosaic augmentation, random brightness, and colour contrast, which creates generalization while training the accurate wheat head detection model.

3.2. Proposed Method

This study aims to design a lightweight network for wheat head detection from field images. The proposed architecture consists of two parts, namely the backbone and detection layer. sample wheat dtection architecture is shown in Figure 1 We have explored the use of modified MobileNetV3 [17] as the backbone, and the feature pyramid structure is utilised in the proposed WheatNet Lite-Neck as a detection layer.

3.2.1 WheatNet-Lite Backbone

We use YOLOV3 [25] as our baseline architecture for object detection. Darknet53 is replaced with the proposed WheatNet-lite backbone. The proposed Wheatnet-Lite backbone use MobileNetV3 as the base network. Mo-

Table 2. Comparison of YOLOV3 and WheatNet-Lite in terms of Parameters

Method	Backbone	Neck	Total Parameter
YOLOV3	37.0 M	25.4 M	62.4 M
WheatNet-Lite	**3.0 M**	**5.2 M**	**8.2 M**

bileNetV3 is based on an inverted residual structure with linear bottlenecks [15]. Depthwise Convolution (DwConv) is modified with Mixed Depthwise Convolution (MDwConv), which splits input channels into groups and apply convolution operation with multiple size kernel to extract multi-scale features of wheat head [31]. MDWConv bottleneck uses a stack of four blocks, including 1 x 1 pointwise convolution, 3 x 3, 5 x 5 filters in mixed depth-wise convolution, squeeze and excitation, and 1 x 1 pointwise linear convolution. Furthermore, a residual skip connection is inserted in the bottleneck, as shown in Figure 5 (a) and Figure 5 (b). The MDWConv bottleneck uses ReLu or h-swish activation function to introduce non-linearity in a network. The complete architecture of the proposed lightweight backbone is shown in Figure 5. The proposed backbone includes one 3 x 3 standard convolution layer followed by 13 MDWConv layers, and finally, one 1 x 1 pointwise convolution. Out of 13 MDWConv bottlenecks, eight bottlenecks are with SE block, and the rest are without SE block. Further, at the end of the backbone, Modified Spatial Pyramid Pooling (MSPP) is utilized with 3×3, 5×5, and 7×7 pool size to extract multi spatial features [13] as shown in Figure 4. As a result, the WheatNet-Lite backbone provides 12 times parameter reduction compared to the standard YOLOV3 backbone, from 37.0 M to 3.0 M, as shown in Table 2. In addition, the intermediate feature maps from the backbone F1, F2, and F3 are extracted at different scales and passed to the proposed WheatNet-Lite detection layer for final wheat head detection, as shown in Figure 6.

3.2.2 WheatNet Lite-Neck

WheatNet-Lite Neck is used for efficient multi-spatial feature extraction and bounding box detection. WheatNet-Lite Neck uses Feature Pyramid Network (FPN) for strengthening the features of the WheatNet-Lite backbone. We replaced standard convolution with depth-wise convolution for parameter reduction. The WheatNet-Lite Neck network provides eight times parameter reduction as compared to standard YOLOV3, from 25.4 M to 5.2 M, as shown in Table 2. The detailed layer-wise structure is shown in Figure 6. The multi-scale features F1 (13 x 13 x 960), F2 (26 x 26 x 672) and F3 (52 x 52 x 240) were extracted from WheatNet-Lite backbone and further given to WheatNet-Lite Neck network. The structure uses 1 x 1 point-wise convolution and depth-wise convolution with upsampling layer, and three multi-scale detection layers are produced for wheat head detection. The output dimensions of WheatNet-Lite are 13 x 13 x 18 for large wheat heads detection, 26 x 26 x 18 for medium wheat heads detection, and 52 x 52 x 18 for small wheat heads detection. The final network includes

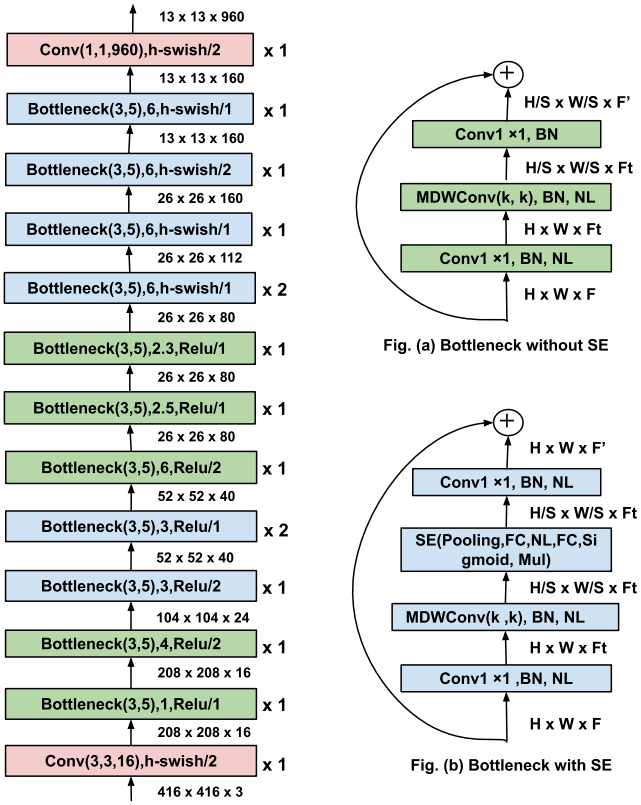

Figure 5. Proposed WheatNet-Lite backbone on left side with Bottleneck layers expressed as Bottleneck(kernel size, expansion factor, Non-linearity / Stride) Fig.(a) shows bottleneck where MDWConv is with 3×3 and 5×5 filters denoted as (k,k), Conv is regular convolution, BN is batch norm, NL is Non linearity Fig.(b) shows bottleneck with squeeze and excitation block. $H \times W \times F$ denotes tensor shape (height, width, depth), t is expansion factor and x /1/2 denotes number of repetition of block.

nine DWConv, seventeen 1 x 1 point-wise convolutions with only 5.2 M parameters in Neck shown in Table 2. It leverages the benefits of FPN and DWConv for more efficient and accurate feature extraction, thereby giving an accurate wheat head detection.

3.2.3 Bounding Box Prediction

After extracting wheat head features, the proposed WheatNet-lite network performs wheat head bounding box detection based on the regression. We use the same loss function as in YOLOV3. Due to variation in the sizes of bounding boxes, we collected three scaled outputs from the WheatNet-Lite Neck 13 x 13 x 18, 26 x 26 x 18, and 52 x 52 x 18. We adopted K-Means++ (clustering technique) to

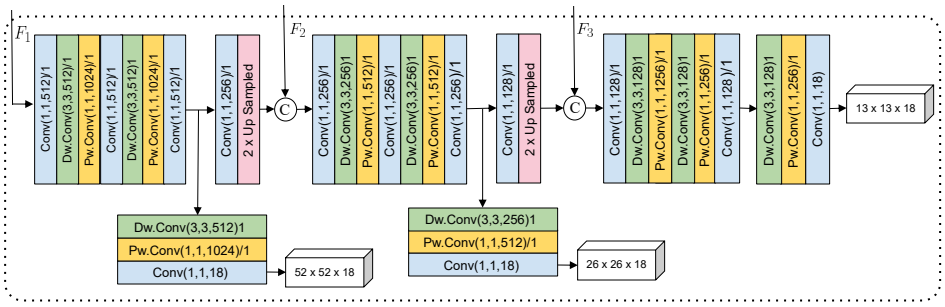

Figure 6. Proposed WheatNet-Lite Detection Layer.

Table 3. Comparison with state of the art methods for GWHD-Train Dataset

Methods	mAP @0.5	Parameter
YOLOV3 [25]	86.9	62.4 M
YOLOV4 [1]	88.53	63.9 M
Zhu et al. [36]	53.64	-
Yang et al.[34]	56.46	62.4 M
Madec et al.[20]	71.32	136.8 M
He et al. [14]	77.68	63.9 M
WheatNet-Lite	**91.32**	**8.2 M**

obtain prior anchor boxes in feature maps by setting three anchor boxes for each scale to collect nine anchor boxes for each cell. The purpose of the K-Means++ algorithm is to cluster the most prior anchor boxes with higher Intersection Over Union(IOU) values with the ground truth-bound boxes. In the proposed network, there are nine anchor boxes due to three scales of feature maps. For each bounding box, there are five expected values, including the bounding box's centre point coordinates(tx, ty), width(tw), height(th), objectness score(Pc), and class probabilities(p1..pi). Finally, it calculates each class's confidence score by multiplying the probabilities of the conditional class and individual box confidence score. For each bounding box score, the corresponding scores of each class are determined as follows:

$$\begin{aligned} Score_n &= P(C_n|Wheathead) \times Pr(Wheathead) \\ &\quad \times IOU_{pred}^{truth} \\ &= P(C_n|Wheathead) \times Conf \end{aligned} \tag{1}$$

4. Experimental Results

In this section, we provide a detailed description of our experimental setup. First, the Evaluation Metrics are mentioned, and then the training strategies for different methods are mentioned. Finally, the performance of the proposed method is evaluated and compared against the state-of-the-art method from both quantitative and qualitative standpoints.

The WheatNet-Lite is implemented using Keras with Tensorflow in the backend as a deep learning framework for training purposes [5]. The experimental environment is Ubuntu 16.04 operating system, Intel (R) Xeon (R) CPU, 2.54 GHz Processor, 16 GB running memory (RAM), NVIDIA GPU Tesla P100 with loss function same as YOLOV3 [25]. The proposed model training is carried out with a batch size of 16 with Adam Optimizer. The learning rate is initially set to 0.001. Models often benefit from reducing the learning rate by monitoring validation loss, and when validation loss remains constant for consecutive five epochs learning rate is reduced by a factor of 0.3. The complete network is trained for 250 epochs. The end-to-end WheatNet-Lite model is trained for each dataset GWHD-Train [4], ACID [22] and SPIKE [12] dataset with same hyper-parameters and same environmental conditions for a fair comparison. Each dataset is split into training, validation, and testing 60%, 20%, and 20%, respectively. Input image size is fixed at $416 \times 416 \times 3$. The model is trained with the training dataset and evaluate it performance in the test dataset. The results of GWHD-Train, ACID, and SPIKE datasets are given in Table 5 and Table 3.

5. Results and Discussion

Mean Average Precision (mAP) is used as an object detection evaluation metric. Experiments started with the selection of feature extractor as the backbone. The Lightweight networks such as EfficientNet [30], ShuffleNet [35], ShuffleNetv2 [19], MobileNet [16], MobileNet-v2 [29] and MobileNet-v3 [15] have great performance on ImageNet dataset which indicates strength of each network. In order to find out which network is best, we conducted experiments using these networks as a backbone of the proposed WheatNet-Lite Network. MobileNet-v3 has the best accuracy in wheat heads detection as the backbone. Thus MobileNet-v3 is chosen as the base backbone network for feature extraction. For the proposed WheatNet-Lite backbone, all DWConv convolutions have been replaced with MDWConv in a standard MobileNet-V3.

The use of MDWConv in inverted residual structure and DWconv in feature pyramid structure is utilized to extract multi-scale features of wheat heads. The proposed method achieved 91.32 mAP@0.5, 76.85 mAP@0.5 and 86.10

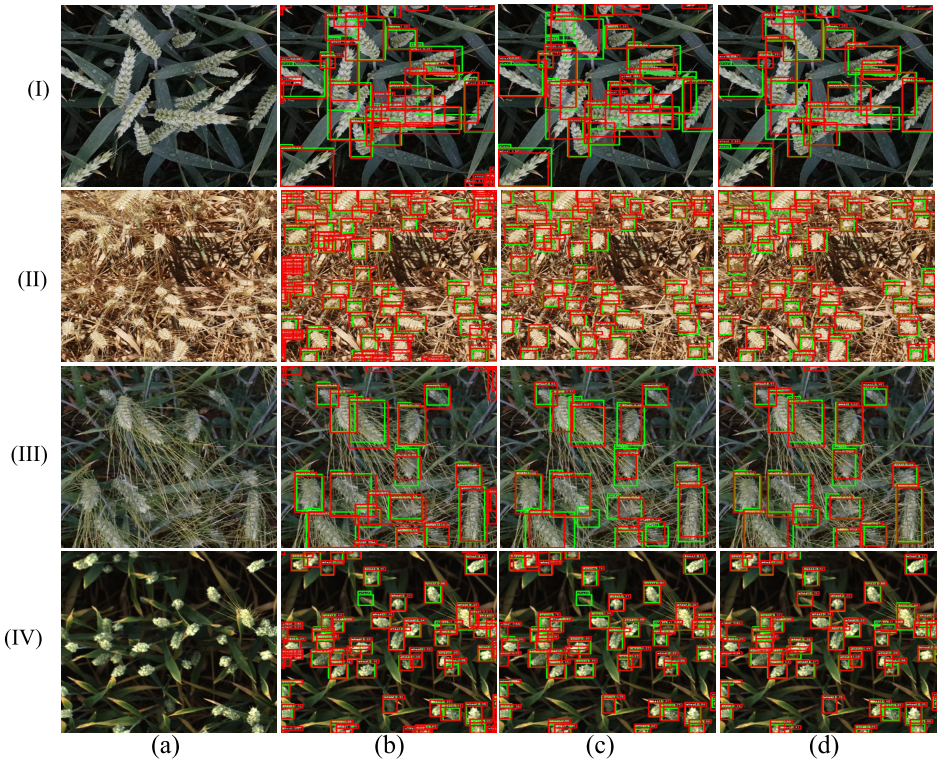

Figure 7. Results for GWHD-Train dataset (a) Original RGB image, (b) Fast-RCNN, (c) YOLOV3, (d) WheatNet-Lite and challenges as (I) overlapping wheatheads, (II) growth stage, (III) Blur image due to wind and (IV) brightness variation. (green box represents ground truth and red box represents predictions)

Table 4. comparison with state of the art methods for GWHD-Train Dataset

Methods	Backbone	Neck	Precision(%)	Recall(%)	mAP @0.5	Parameter	GFLOPS	Time
YOLOV2	DarkNet19	-	70.17	75.13	71.32	50.6 M	53.10 G	28ms
YOLOV3	Darknet53	FPN	89.26	91.12	86.9	62.4 M	65.52 G	30ms
YOLOV4	CSPDarknet53	PAN	90.12	93.12	88.53	63.9 M	67.07 G	38ms
WheatNet-Lite	**Proposed**	**Proposed**	**91.32**	**94.32**	**91.32**	**8.2 M**	**5.2 G**	**15ms**

Table 5. Comparison of proposed WheatNet-Lite on different Datasets

Dataset	mAP @0.5	Precision	Recall	Parameters
GWHD-Train	91.32	89.85	92.45	8.2 M
ACID	76.85	80.69	85.37	8.2 M
SPIKE	86.10	81.56	90.32	8.2 M

mAP@0.5 for GWHD-Train, ACID and SPIKE datasets with only 8.2 M parameters, as shown in Table 5. Additionally, The quantitative results comparison with state-of-the-art methods on the GWHD-Train dataset is shown in Table 3. The qualitative results comparison of the proposed WheatNet-Lite with fast-RCNN and YOLOV3 shown in Figure 7. columns in figure 7(a) represents input original image 7 (b), 7 (c) and 7 (d) represents results of Fast-RCNN, YOLOV3 and WheatNet-Lite respectively. Proposed methods overcoming the challenges of wheat head detection are shown in 7 . such as 7(I) overlapping wheat heads,, 7(II) growth stage, 7(III) Blur image due to wind and 7(IV) brightness variation. The qualitative results for

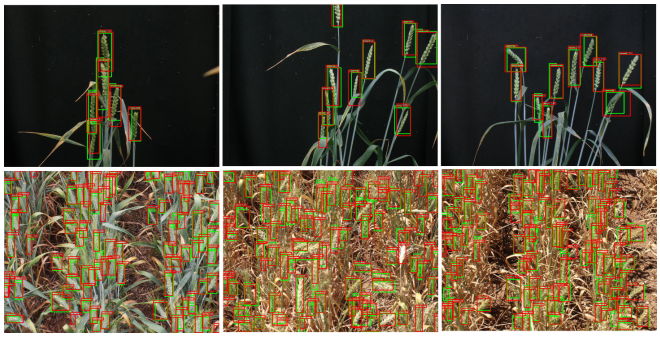

Figure 8. WheatNet-Lite results on ACID and SPIKE dataset green and red box represent ground truth and prediction respectively (first and second row for ACID and SPIKE dataset respectively)

the ACID, and SPIKE datasets are shown in Figure 8. Even though for GWHD-Train and SPIKE dataset, ground truth-bounding boxes are missing for some wheat heads, the proposed method predicts correct bounding box for wheat

Table 6. Ablation Study of WheatNet-Lite on GWHD-Train Dataset

Methods	Proposed backbone	MSPP	Lite Neck	DWConv	MDWConv	Map @0.5	Parameter	GFLOPS
YOLOV3						86.9	62.4 M	65.52 G
WheatNet-Lite-Mod1	✓					87.53	28.4 M	29.82 G
WheatNet-Lite-Mod2	✓	✓				88.78	28.4 M	30.82 G
WheatNet-Lite-Mod3	✓	✓	✓			89.53	8.2 M	8.61 G
WheatNet-Lite-Mod4	✓	✓	✓	✓		90.5	8.2 M	5.16 G
WheatNet-Lite	✓	✓	✓		✓	**91.32**	**8.2 M**	**5.2 G**

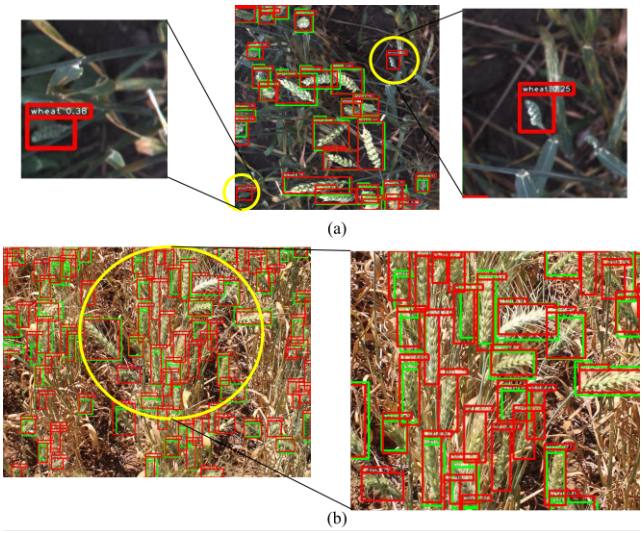

Figure 9. WheatNet-Lite performs very well even without ground truth bounding boxes examples from (a) GWHD-Train and (b) SPIKE dataset.(green and red box represent ground truth and prediction respectively)

heads as shown in Figure 9. Even if the proposed method predicts correct wheat head, because of its ground truth is not available, it is considered false positive. Due to this, mAP score decreases. Adding up ground truth for these wheat heads in both the dataset will increase the results.

5.1. Ablation Study

In the ablation study the proposed light weight network is selected. Mixed depthwise convolution, MSPP [13] and Depthwise convolution is effectively used in backbone and Lite-Neck respectively.

YOLOV3 is selected as a baseline architecture, backbone DarkNet53 is replaced with modified MobileNetV3 as Modification 1 and added MSPP as modification 2 shows an increment in results. Further, DWConv utilized with feature pyramid structure in WheatNet-Lite Neck as Modification 3. Finally, MDWConv use in modified MobileNetV3 as Modification 4 gives more accurate results than the other state of the art methods shown in Table 6. Observations about the Mixed depthwise in the inverted residual block is multi-scale features are extracted for more accurate wheat head detection, as shown in Table 6. Departing from standard convolution, we have utilized mixed depthwise convolution and depthwise convolution for efficient model designing to reduce parameters. Furthermore, we have demonstrated the effectiveness of the proposed method for multi-scale feature extraction for accurate wheat head detection by overcoming challenges such as occlusion of the wheat head, genotype, canopy and growth stage variation, and image blur due to wind.

6. Conclusion

This paper presents a lightweight WheatNet-Lite convolutional neural network-based method for wheat head detection. The proposed method consists of a WheatNet-Lite backbone network using Mixed-Depthwise Convolution (MDWConv) with inverted residual block. The Modified Spatial Pyramid Pooling (MSPP) layer is effectively used for multi-scale feature extraction in the backbone. Finally, the DWConv is used in the feature pyramid structure with WheatNet-Lite neck for bounding box prediction. The proposed method outperforms the existing state-of-the-art methods with the reduced parameters (eight times) compared to YOLOV3. The achieved results with mean average precision (MAP) 93.6 MAP@0.3, 91.32 MAP@0.5 with only 8.2 M parameters on GWHD-Train dataset. Also, the method evaluates the new proposed ACID dataset with bounding box annotations with inference time as 10 ms. The extensive experiments carried out on GWHD-Train, SPIKE and ACID datasets show 91.32 mAP@0.5, 86.10 mAP@0.5 and 76.85 mAP@0.5, respectively. Thus, we have shown empirically that our approach is much more suited for real-time wheat head detection. In future, the proposed approach can effectively make an impact on real-time wheat head detection.

Acknowledgements

This publication is outcome of research and development work under the NDF scheme governed by AICTE, New Delhi, India and funded by MHRD, India.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[2] P.W.D. Charles. Project title. https://github.com/tzutalin/labelImg, 2013.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[4] Etienne David, Simon Madec, Pouria Sadeghi-Tehran, Helge Aasen, Bangyou Zheng, Shouyang Liu, Norbert Kirchgessner, Goro Ishikawa, Koichi Nagasawa, Minhajul A Badhon, 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. *Plant Phenomics*, 2020, 2020.

[5] Joshua V Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A Saurous. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

[6] FAOSTAT. Available online. https://www.fao.org/faostat/zh/#data/QC, accessed on 29 December 2020.

[7] Jose A Fernandez-Gallego, Ma Buchaillot, Nieves Aparicio Gutiérrez, María Teresa Nieto-Taladriz, José Luis Araus, Shawn C Kefauver, et al. Automatic wheat ear counting using thermal imagery. *Remote Sensing*, 11(7):751, 2019.

[8] Jonathon A Gibbs, Alexandra J Burgess, Michael P Pound, Tony P Pridmore, and Erik H Murchie. Recovering wind-induced plant motion in dense field environments via deep learning and multiple object tracking. *Plant physiology*, 181(1):28–42, 2019.

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[10] Bo Gong, Daji Ergu, Ying Cai, and Bo Ma. A method for wheat head detection based on yolov4. 2020.

[11] Bo Gong, Daji Ergu, Ying Cai, and Bo Ma. Real-time detection for wheat head applying deep neural network. *Sensors*, 21(1):191, 2021.

[12] Md Mehedi Hasan, Joshua P Chopin, Hamid Laga, and Stanley J Miklavcic. Detection and analysis of wheat spikes using convolutional neural networks. *Plant Methods*, 14(1):1–13, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[14] Ming-Xiang He, Peng Hao, and You-Zhi Xin. A robust method for wheatear detection using uav in natural scenes. *IEEE Access*, 8:189043–189053, 2020.

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

[16] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[17] Saeed Khaki, Nima Safaei, Hieu Pham, and Lizhi Wang. Wheatnet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting. *arXiv preprint arXiv:2103.09408*, 2021.

[18] Maria Lusser, Claudia Parisi, Damien Plan, Emilio Rodríguez-Cerezo, et al. *New plant breeding techniques: state-of-the-art and prospects for commercial development*. Publications Office of the European Union Luxembourg, 2011.

[19] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[20] Simon Madec, Xiuliang Jin, Hao Lu, Benoit De Solan, Shouyang Liu, Florent Duyme, Emmanuelle Heritier, and Frederic Baret. Ear density estimation from high resolution rgb imagery using deep learning technique. *Agricultural and forest meteorology*, 264:225–234, 2019.

[21] Narendra Narisetti, Kerstin Neumann, Marion S Röder, and Evgeny Gladilin. Automated spike detection in diverse european wheat plants using textural features and the frangi filter in 2d greenhouse images. *Frontiers in Plant Science*, 11:666, 2020.

[22] Michael P Pound, Jonathan A Atkinson, Darren M Wells, Tony P Pridmore, and Andrew P French. Deep learning for multi-task plant phenotyping. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2055–2063, 2017.

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[24] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[28] Pouria Sadeghi-Tehran, Nicolas Virlet, Eva M Ampe, Piet Reyns, and Malcolm J Hawkesford. Deepcount: in-field automatic quantification of wheat spikes using simple linear iterative clustering and deep convolutional neural networks. *Frontiers in plant science*, 10:1176, 2019.

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[30] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[31] Mingxing Tan and Quoc V Le. Mixconv: Mixed depthwise convolutional kernels. *arXiv preprint arXiv:1907.09595*, 2019.

[32] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[33] Kaaviya Velumani, Sander Oude Elberink, Michael Ying Yang, and Frédéric Baret. Wheat ear detection in plots by segmenting mobile laser scanner data. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4, 2017.

[34] Yvhang Yang, Xi Huang, Liangben Cao, Lihong Chen, and Kailiang Huang. Field wheat ears count based on yolov3. In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pages 444–448. IEEE, 2019.

[35] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[36] Yanjun Zhu, Zhiguo Cao, Hao Lu, Yanan Li, and Yang Xiao. In-field automatic observation of wheat heading stage using computer vision. *Biosystems Engineering*, 143:28–41, 2016.