Computers and Electronics in Agriculture 211 (2023) 108006

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Deformable convolution and coordinate attention for fast cattle detection

Wenjie Yang [a], Jiachun Wu [c], Jinlai Zhang [b,c,*], Kai Gao [b], Ronghua Du [b], Zhuo Wu [e], Eksan Firkat [e], Dingwen Li [d]

[a] College of Mathematics and Statistics, Hu'nan University of Finance and Economics, Changsha, 410205, Hunan, China
[b] College of Automotive and Mechanical Engineering, Changsha University of Science and Technology, Changsha, 410114, Hunan, China
[c] College of Mechanical Engineering, Guangxi University, Nanning, 530004, Guangxi, China
[d] Aecc South Industry Company Limited, Zhuzhou, 412000, Hunan, China
[e] School of Information Science and Engineering, Xinjiang University, Urumqi, 830000, Xinjiang, China

ARTICLE INFO

Keywords:
YOLOv8
Computer vision
Precision agriculture

ABSTRACT

Cattle detection is an important task in precision livestock farming, but it remains challenging due to the varying appearance and poses of cattle in different scenarios. In this paper, we propose a novel approach for fast cattle detection using deformable convolution and coordinate attention within YOLOv8, a SOTA object detection model. Our proposed method enhances the YOLOv8 architecture by introducing deformable convolution to capture more fine-grained spatial information and coordinate attention to emphasize important features in the detection process. We evaluate our method on a cattle dataset collected in a cattle farm and achieve superior performance compared to the baseline YOLOv8 and several SOTA object detection models. Specifically, our approach achieves a mean average precision (mAP) of 72.9% at 62.5 frames per second (FPS), which demonstrates its effectiveness and efficiency for fast cattle detection. By deploying our method on the farm's monitoring computer, our proposed approach has the potential to facilitate the development of automated cattle monitoring systems for improving animal welfare and farm management.

1. Introduction

Cattle farming plays a vital role in the global economy, providing food, leather, and other products to humans (Qiao et al., 2023). As the demand for animal products continues to rise, there is a growing need for efficient and effective cattle management systems to improve animal welfare and farm productivity. Precision livestock farming (Banhazi et al., 2012) (PLF) has emerged as a promising approach to address this challenge by using sensor technologies, machine learning (Qiao et al., 2022), and other advanced techniques to monitor, analyze, and optimize animal health, behavior, and performance. Cattle detection is a key task in PLF, which aims to automatically locate and recognize individual cattle in real-time based on their visual appearance and behavior.

Object detection is a fundamental task in computer vision (Zou et al., 2023), which aims to locate and classify objects of interest in an image or video. With the recent advances in deep learning (Weng et al., 2022), object detection has achieved remarkable progress and has become a crucial component in many real-world applications, such as agricultural robots, security surveillance, and medical diagnosis. YOLO (Redmon et al., 2016) (You Only Look Once) is a popular object detection framework that can achieve real-time performance with high accuracy. YOLOv8 is the latest version of YOLO, which utilizes a CSP-Darknet (Glenn, 2022) backbone. Despite the impressive performance of YOLOv8, cattle detection remains a challenging task due to several factors. First, cattle have various appearances and poses, which make them difficult to distinguish from the background and other animals. Second, cattle can move quickly and unpredictably, which poses a challenge for real-time detection. Third, cattle may be occluded by other objects or animals, which can result in false negatives or false positives. Moreover, YOLOv8 shows a poor performance under common corruption. Therefore, there is a need to develop novel methods that can improve the accuracy and efficiency of cattle detection in PLF. In recent years, deformable convolution (Dai et al., 2017) and attention mechanism (Qiao et al., 2023) have shown great potential in enhancing the performance of image classification. Deformable convolution can adaptively adjust the receptive fields of convolutional layers to capture more fine-grained spatial information, it can increase the receptive field without increasing the number of parameters and helps in better capturing the spatial information which is useful to distinguish cattle from their background. In addition, the attention mechanism can selectively emphasize important features and suppress irrelevant

* College of Automotive and Mechanical Engineering, Changsha University of Science and Technology, Changsha 410114, China.
E-mail address: 228434973@qq.com (J. Zhang).

https://doi.org/10.1016/j.compag.2023.108006
Received 8 March 2023; Received in revised form 16 June 2023; Accepted 17 June 2023
Available online 4 July 2023

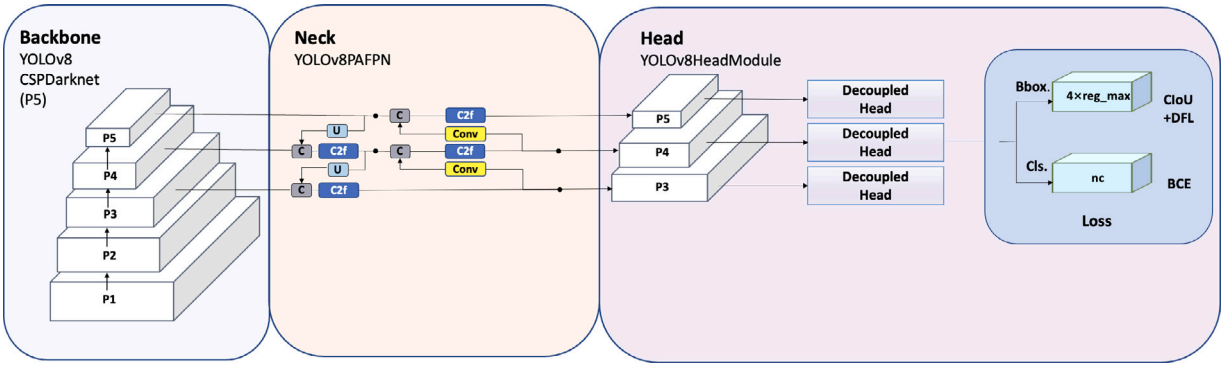

Fig. 1. Detailed architecture of YOLOV8. "C" denotes the Concat operation, "U" denotes the upsampling operation, and "C2f" denotes the C2f module in YOLOv8.

ones. Several studies have applied attention mechanism (Wang et al., 2022c,b) to object detection, and achieved state-of-the-art results on various agricultural applications.

In this paper, we propose a novel approach for fast cattle detection using deformable convolution and coordinate attention within YOLOv8. Our proposed method enhances the YOLOv8 architecture by introducing deformable convolution to capture more fine-grained spatial information and coordinate attention to emphasize important features in the detection process, we call this method DCA-YOLOv8 for short. Specifically, we add deformable convolution to the backbone and introduce coordinate attention to the detection head of YOLOv8. We evaluate our method on a cattle dataset established using surveillance cameras on a farm and compare it with the baseline YOLOv8 and several state-of-the-art (SOTA) object detection models. Specifically, DCA-YOLOv8s surpassed SSD, RetinaNet, and CornerNet in mAP indicators by 13, 4.9, and 15.1, respectively. At the same time, the inference speed is 156.25 times that of CornerNet. Our proposed approach achieves superior performance in terms of accuracy and efficiency, which demonstrates its effectiveness in cattle detection. The contributions of this paper are as follows:

- We propose a novel approach for fast cattle detection using deformable convolution and coordinate attention within YOLOv8.
- We evaluate our method on a cattle dataset collected in a cattle farm and compare it with the baseline YOLOv8 and several SOTA object detection models.
- We demonstrate that our proposed approach achieves superior performance in terms of accuracy and efficiency, which shows its potential for automated cattle monitoring systems in PLF.

The rest of the paper is organized as follows. Section 2 describes the proposed approach in detail, including the network architecture. Section 3 describes the experimental setup, including the collection of the dataset, data augmentation used during training and implementation details. Section 4 presents the experimental results and analysis, followed by a discussion in Section 5. Finally, Section 6 concludes the paper and summarizes the contributions.

2. Proposed approach

In this paper, we propose a novel approach for fast cattle detection using deformable convolution and coordinate attention. Our proposed method enhances the YOLOv8 architecture by introducing **D**eformable **C**onvolution to capture more fine-grained spatial information and coordinate **A**ttention to emphasize important features in the detection process, we call this method **DCA-YOLOv8** for short.

2.1. Overview of YOLOv8

The YOLOv8 (Glenn, 2023) architecture is a variant of the YOLO (You Only Look Once) (Redmon et al., 2016) family of object detection models that uses a single neural network to predict the bounding boxes and labels of objects in an image. The YOLOv8 architecture consists of several key components, including a backbone network, neck network, and detection head. The backbone is used to extract feature maps from the input image, while the neck network and head network are used for predicting the bounding boxes and labels of objects in the feature maps. The overview of YOLOv8 is shown in Fig. 1. Following, we will introduce the backbone network and detection head in detail.

Backbone and Neck. The backbone and neck parts of the model may have been inspired by the design of YOLOv7 ELAN (Wang et al., 2022a), in which the C3 structure of YOLOv5 (Glenn, 2022) was replaced with the C2f structure that has a richer gradient flow, and the number of channels was adjusted for different scale models, the details of the C2f module can refer to Glenn (2023). This is a careful adjustment of the model structure, rather than blindly applying a set of parameters to all models, which significantly improves the model performance (Glenn, 2023). Moreover, the neck network is to bridge the gap between the feature representation output from the backbone network and the head network's prediction.

Detection head. Compared with YOLOv5, the detection head has undergone major changes and has been replaced with the currently popular decoupled head structure, which separates the classification and detection heads. As shown in Figs. 2 and 3,

Loss. The YOLOv8 architecture uses a loss function that is a combination of the localization loss and classification loss, as described in the previous section. The total loss is defined as:

$$L = \lambda_{loc} L_{loc} + \lambda_{cls} L_{cls}, \tag{1}$$

where λ_{loc} and λ_{cls} are hyperparameters that control the relative importance of the localization loss and classification loss.

2.2. The DCA-YOLOv8 detection model

Our proposed network architecture is shown in Fig. 4. Firstly, to capture the scale variation of cattle caused by their distances to the camera in the farm, we introduced deformable convolution (Dai et al., 2017). Specifically, we replaced all the convolutions in the C2f module of the YOLOv8 backbone network with deformable convolutions to improve the model's performance under scale variations of cattle. Secondly, to better distinguish between small cattles with similar appearances but different spatial locations, we introduce coordinate attention (Hou et al., 2021) module before the small detection head. Next, we will explain these two modules in detail.

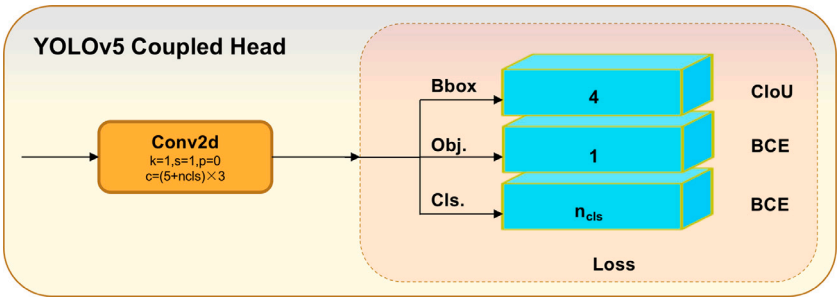

Fig. 2. YOLOV5 head.

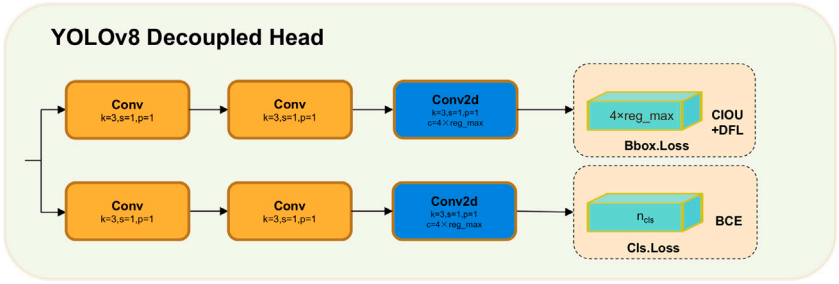

Fig. 3. YOLOV8 head.

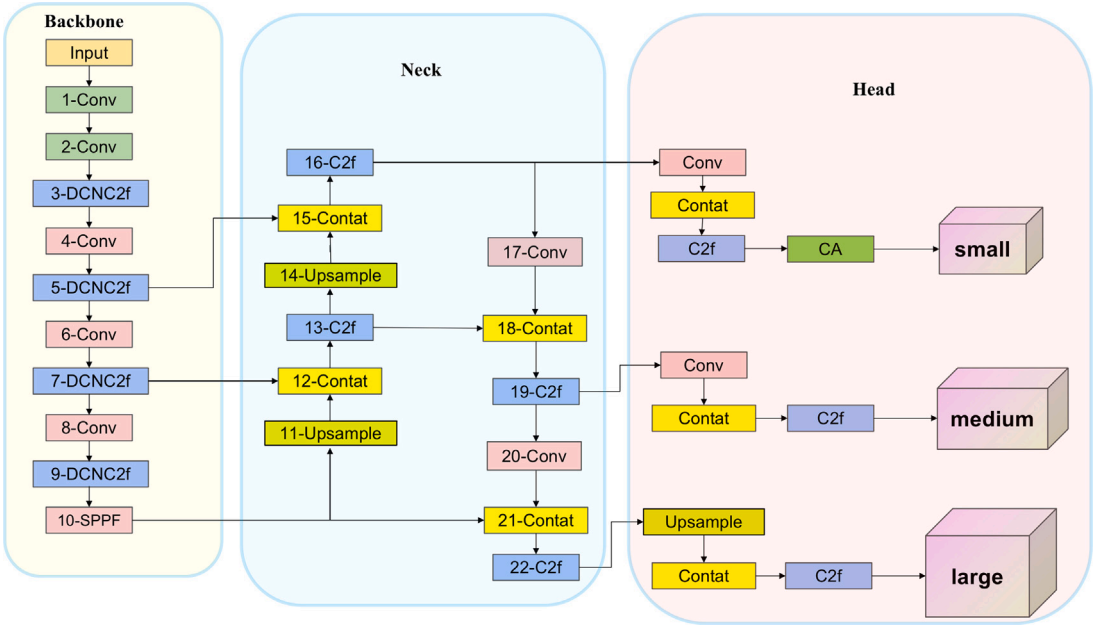

Fig. 4. Detailed architecture of the proposed DCA-YOLOv8. *small, medium, large* denote the detection head with different sizes.

2.2.1. Deformable convolution module

Deformable Convolutional Networks (DCN) (Dai et al., 2017) are a type of convolutional neural network that was designed to address the limitations of traditional convolutional neural networks (Zhang et al., 2020). In a traditional convolutional neural network, the convolutional operation is performed using a fixed grid of sampling points, also known as the receptive field (Wei et al., 2022). However, this fixed grid may not be able to capture the complex spatial variations of objects in the image, especially when the object is distorted or small.

To overcome this limitation, DCN introduces a new module called deformable convolution, the detailed structure of deformable convolutional is shown in Fig. 5. In a deformable convolution, the traditional sampling grid is replaced with a learnable sampling grid that can be dynamically adjusted according to the object's shape and deformation. This allows the model to capture more fine-grained spatial information and improve its detection accuracy. Let F be the input feature map and W be the weight tensor of the deformable convolution. The deformable convolution operation can be formulated as follows:

$$Y_{i,j} = \sum_{p,q} F_{i+p,j+q} \times W_{p,q,k} \tag{2}$$

where $Y_{i,j}$ is the output feature map, and p and q are the offset values that determine the sampling location of the input feature map. The offset values are learned during the training process and can be expressed as:

$$p = p_0 + \Delta p, \ q = q_0 + \Delta q \tag{3}$$

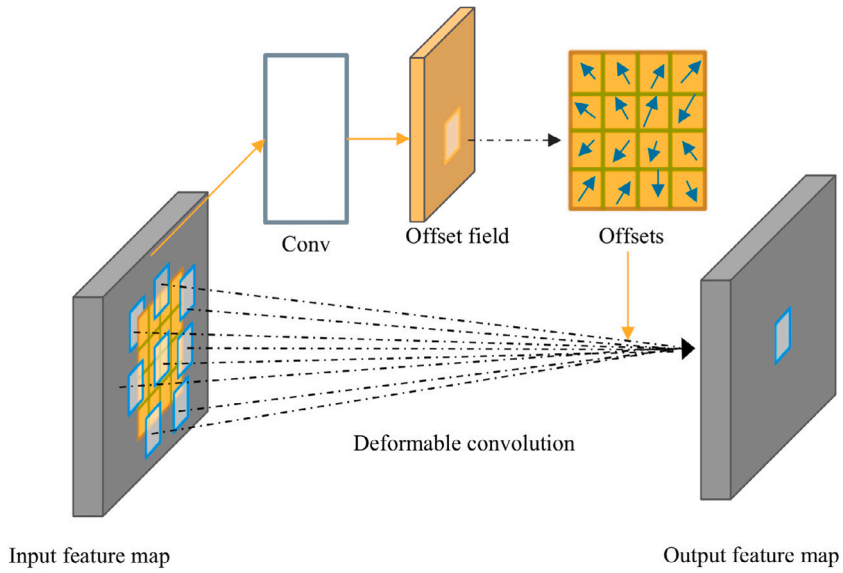

Fig. 5. Deformable convolution.

where (p_0, q_0) is the center location of the receptive field (He et al., 2016), and $(\Delta p, \Delta q)$ is the learned offset value. The offset value is computed based on the input feature map F and the output feature map Y, which allows the model to learn the most appropriate sampling location for each point.

Moreover, DCN introduces a modulated deformable convolution to further enhance the flexibility of the sampling grid. The modulated deformable convolution adds an additional modulation term to the weight tensor, which is multiplied by the learned offset value. The modulated deformable convolution operation can be formulated as:

$$Y_{i,j} = \sum_{p,q} F_{i+p,j+q} \times W_{p,q,k} \times M_{p,q,k} \tag{4}$$

where $M_{p,q,k}$ is the modulation term, which is computed based on the input feature map F and the output feature map Y. The modulation term allows the model to adjust the weight tensor based on the object's shape and deformation, this is useful for cattle detection. Moreover, many recent works have also empirically verified that DCN can be applied to challenging scenarios such as occlusion, varying appearances and different poses (Chilukuri et al., 2022; Fang et al., 2022).

2.2.2. Coordinate Attention module

Coordinate Attention (Hou et al., 2021) is a recently proposed attention mechanism that is designed to capture the long-range dependencies of features in an image. Unlike traditional attention mechanisms that attend to all spatial locations equally, Coordinate Attention selectively attends to certain spatial locations that are relevant to the task at hand, the detailed structure of deformable convolutional is shown in Fig. 6. The Coordinate Attention operates by modeling the relationships between different spatial locations in an image. It does this by first projecting the input feature map F onto two sets of features Q and K:

$$Q = W_Q \cdot F, \ K = W_K \cdot F, \tag{5}$$

where W_Q and W_K are learnable projection matrices. The feature sets Q and K are then used to compute the attention map A:

$$A_{i,j} = \sum_{u=1}^{H} \sum_{v=1}^{W} \mathrm{softmax}\left(\frac{Q_{i,j} \cdot K_{u,v}}{\sqrt{d}} \right) \cdot V_{u,v}, \tag{6}$$

where H and W are the height and width of the input feature map, respectively, d is the dimensionality of the feature vectors, and V is the value feature map. The attention map A captures the relationships between all spatial locations in the input feature map.

Coordinate Attention further enhances the attention map A by adding a coordinate embedding term E:

$$A_{i,j} = \sum_{u=1}^{H} \sum_{v=1}^{W} \mathrm{softmax}\left(\frac{Q_{i,j} \cdot K_{u,v}}{\sqrt{d}} + E_{i,j,u,v} \right) \cdot V_{u,v}, \tag{7}$$

where $E_{i,j,u,v}$ is the coordinate embedding term that encodes the relative position between the spatial location (i, j) and (u, v). The coordinate embedding term allows the attention mechanism to selectively attend to certain spatial locations that are relevant to cattle.

2.3. Loss function

In the cattle detection tasks, the goal is to predict the bounding boxes and labels of cattle in an image. To train a model for this task, a loss function is typically used to measure the difference between the predicted outputs and the ground truth annotations.

One commonly used loss function in cattle detection is the sum of two terms: localization loss and classification loss. The localization loss is usually calculated using the smooth L1 loss (Girshick, 2015), which is a modification of the L1 loss that is less sensitive to outliers:

$$L_{loc} = \sum_{i \in Pos} \sum_{j \in x,y,w,h} \mathrm{smooth}_{L1}(t_j^{(i)} - \hat{t}_j^{(i)}), \tag{8}$$

where Pos is the set of positive anchor boxes, $t_j^{(i)}$ is the ground truth value of the jth coordinate of the ith anchor box, $\hat{t}j^{(i)}$ is the predicted value of the jth coordinate of the ith anchor box, and $\mathrm{smooth}\,L1$ is the smooth L1 loss function:

$$\mathrm{smooth}_{L1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise.} \end{cases} \tag{9}$$

The classification loss is usually calculated using the cross-entropy loss, which measures the difference between the predicted class probabilities and the ground truth class labels:

$$L_{cls} = -\sum_{i \in Pos \cup Neg} \sum_{c=1}^{C} y_c^{(i)} \log(\hat{y}_c^{(i)}), \tag{10}$$

where Neg is the set of negative anchor boxes, $y_c^{(i)}$ is the ground truth probability of the ith anchor box belonging to class c, $\hat{y}_c^{(i)}$ is the predicted probability of the ith anchor box belonging to class c, and C is the number of classes.

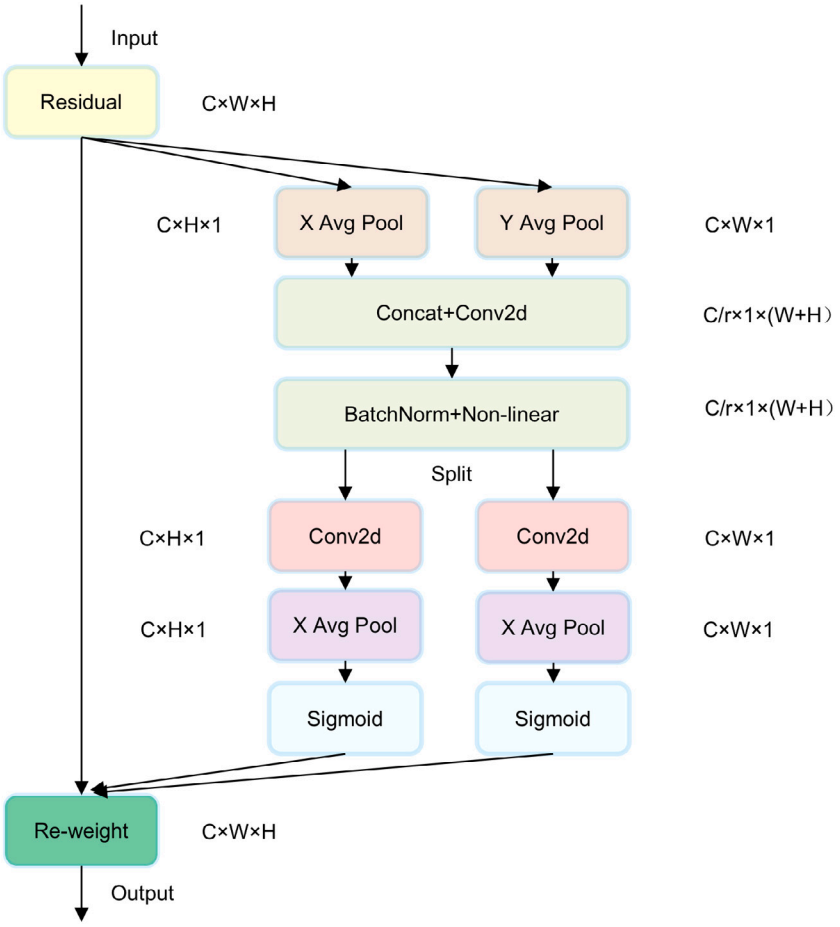

Fig. 6. Detailed structure of coordinate attention module.

Table 1
Details of the dataset.

Class	Number of images	Number of instances
Train	956	9748
Test	240	3542

3. Experimental setup

3.1. Dataset

We select 15 cameras around the Puyuan Farm, Zhuzhou, China. We choose the Hikvision DS-2CD2110F-I as our camera, which has a resolution of 1.3 megapixels (1280 × 960), providing clear and detailed images. In addition, the camera has a fixed 4 mm lens with a wide viewing angle of 70 degrees, and the camera comes with built-in infrared (IR) illuminators that enable it to capture clear images in low-light and dark environments, up to a range of 30 m. The images were captured at different times of day and under different lighting conditions to ensure a diverse range of images, the breed of cattle is Simmental cattle. We collected a dataset of 1196 images from a cattle farm for training and evaluating our DCA-YOLOv8 model for cattle detection.

All the images in the dataset were labeled using the LabelImg (Lin, 2015) tool to create bounding box annotations around the cattle in the images, all bounding boxes were re-confirmed by three experienced farm workers who have worked for several years. The annotations were saved in the YOLO format (Glenn, 2022) and included the coordinates of the bounding boxes, as well as the class labels (i.e., "cattle"). Samples of the collected images are shown in Fig. 7.

The images in the dataset were then divided into two subsets: training and testing sets, as shown in Table 1. The training set contained 956 images and the testing set contained 240 images. The images were split randomly to ensure a diverse range of images in each subset.

3.2. Data augmentation

Data augmentation is an important technique used in computer vision (Zhang et al., 2022) to increase the size and diversity of training data, which can help improve the generalization performance of the detection model. In our study, we used several data augmentation techniques to augment the training set and improve the robustness of our DCA-YOLOv8 model for cattle detection.

We applied random horizontal and vertical flipping, random cropping, random resizing, random rotation, random brightness, and random contrast to the training set, the details of more data augmentations can refer to Su et al. (2021). Fig. 8 shows the images after augmentation for each augmentation. For each image in the training set, we performed each augmentation twice. By applying these techniques, we were able to train a more robust and accurate model for cattle detection.

3.3. Implementation details

We train our proposed method on the collected cattle dataset, which contains 1,196 images with 13,290 cattle instances. We use the same training settings as YOLOv8 (Glenn, 2023), with more data augmentations in the previous section. Following Glenn (2023), we set the input image size to 640 × 640 and the batch size to 32. We use stochastic gradient descent (SGD) with a learning rate of 0.001 and a

Fig. 7. Samples of the data.

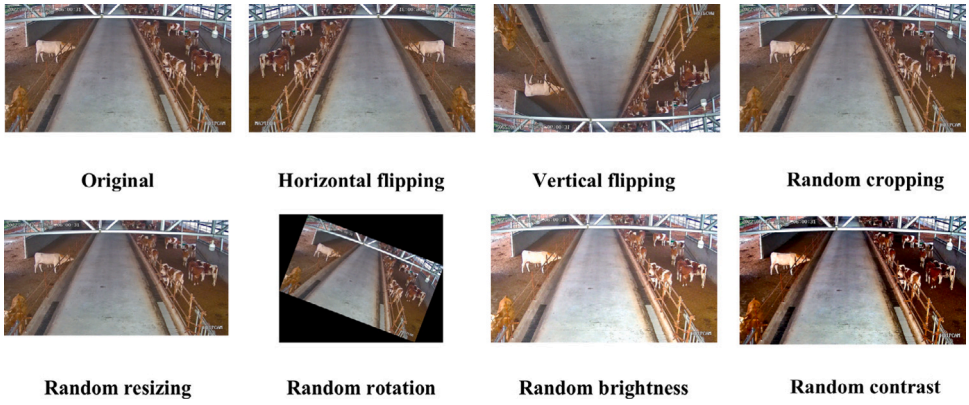

Fig. 8. Samples of the data augmentations.

momentum of 0.9 to optimize the network. We train the network for 100 epochs. All the experiments are implemented on a Linux server with Intel i7 CPU, 64 GB RAM and NVIDIA GTX 3090TI GPU.

4. Experiments

4.1. Evaluation metrics

In order to evaluate the performance of our DCA-YOLOv8 model for cattle detection, we used a variety of evaluation metrics commonly used in object detection tasks.

The first evaluation metric is mean Average Precision (mAP), which is widely used in object detection tasks. mAP measures the average precision of a model at different intersection over union (IoU) thresholds. IoU measures the overlap between the predicted bounding boxes and the ground truth bounding boxes. The formula for mAP can be expressed as:

$$\text{mAP} = \frac{1}{|C|} \sum_{c \in C} \text{AP}(c) \tag{11}$$

where C is the set of object classes, |C| is the number of classes, and AP(c) is the average precision of class c. In this paper, the class is *cattle*.

The second evaluation metric is frames per second (FPS) of YOLOv8 can be calculated using the following formula:

$$\text{FPS} = \frac{1}{\text{inference time per frame}} \tag{12}$$

where the inference time per frame is the amount of time it takes to process a single frame with the YOLOv8 model. In other words, the FPS is the reciprocal of the inference time per frame, which represents the number of frames that can be processed in one second.

4.2. Baseline model comparison

In this section, we compare the performance of our DCA-YOLOv8 model with the baseline YOLOv8 model, which was trained using the same dataset but without deformable convolution (DCN) and coordinate attention (CA).

Table 2
Comparison of baseline models.

DCN	CA	Model size (MB)	FLOPs (G)	mAP@0.5 (%)
		21.45	28.6	72.0
✓		21.56	24.9	72.1
	✓	21.50	28.7	72.3
✓	✓	22.12	**24.9**	**72.9**

Table 2 shows the performance comparison of the two models on the test set. As can be seen, our improved model achieves significantly better performance than the baseline model in terms of mAP and FLOPs. Specifically, our model improves the mAP by 0.9%, and decreased the FLOPs from 28.6 to 24.9.

The results of the confusion matrix for DCA-YOLOv8s and YOLOv8s are shown in Fig. 9. The details of the calculation of confusion matrix can refer to the code open-sourced by (Glenn, 2023). As shown in the figure, the DCA-YOLOv8s performed slightly better in identifying cattle objects, with 76% of cattle objects predicted correctly compared to YOLOv8s' 72%. However, both models still struggled to accurately identify all cattle objects, with a non-negligible proportion being misclassified as background.

Furthermore, we analyzed the changes in box loss, class loss, and mAP during the training process. As shown in Fig. 10, we observed that DCA-YOLOv8s had a faster decrease in training loss compared to YOLOv8s, with similar box loss and class loss. This suggests that DCA-YOLOv8s outperforms YOLOv8s. The change in mAP also indicates that DCA-YOLOv8s significantly outperforms YOLOv8s during training.

Overall, these results demonstrate the effectiveness of our proposed improvements to the YOLOv8 model for cattle detection, and highlight the importance of incorporating deformable convolution and coordinate attention into object detection models for improved performance.

4.3. Comparison with state-of-the-art detection models

In this section, we compare the performance of our DCA-YOLOv8s model with state-of-the-art object detection models for cattle detection.

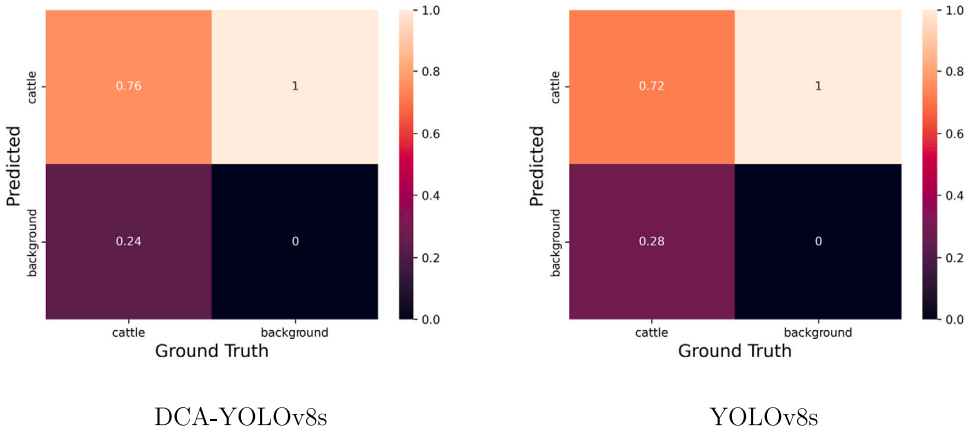

Fig. 9. Confusion matrix of DCA-YOLOv8s and YOLOv8s.

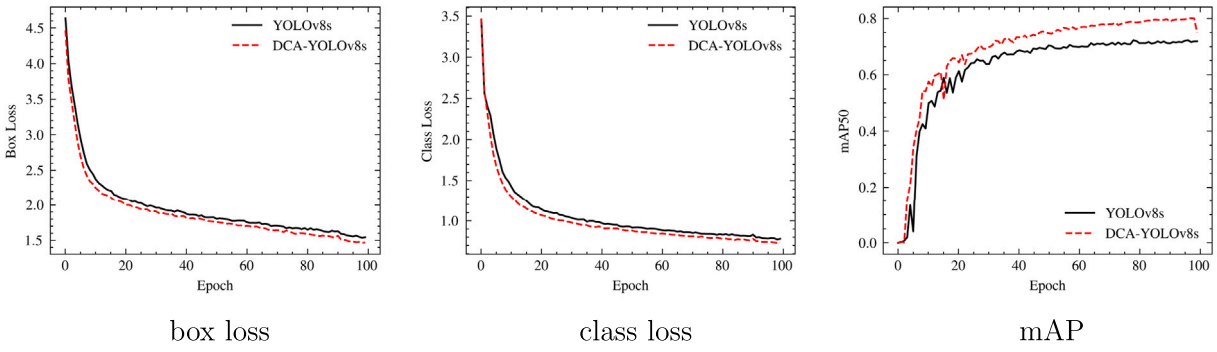

Fig. 10. Analysis of the training process.

Table 3
Performance comparison of the SOTA models.

Model	mAP	FPS	FLOPs	Params
Faster R-CNN (Ren et al., 2015)	**76.9**	4.5	206.6 G	41.1 M
SSD (Liu et al., 2016)	59.9	10.8	342.7 G	23.8 M
RetinaNet (Lin et al., 2017)	68.0	4.9	204.36 G	36.1 M
Cascade R-CNN (Cai and Vasconcelos, 2019)	73.7	3.7	234.46 G	68.9 M
CornerNet (Law and Deng, 2018)	57.8	0.4	1765 G	200.9 M
YOLOv5s (Glenn, 2022)	70.9	46.3	23.8 G	9.1 M
DCA-YOLOv8s (ours)	72.9	**62.5**	24.9 G	11.5 M

Table 4
Comparison of different positions of DCN.

1	2	3	4	Model size (MB)	FLOPs (G)	mAP@0.5 (%)
✓				21.48	28.5	71.2
✓	✓			21.60	27.4	71.4
✓	✓	✓		21.84	25.9	71.9
✓	✓	✓	✓	**22.06**	**24.9**	**72.1**

The Faster R-CNN, SSD, RetinaNet, Cascade R-CNN and CornerNet are implemented using mmdetection toolbox (Chen et al., 2019), and the DCA-YOLOv8s and YOLOv5s are implemented using YOLOv8 (Glenn, 2023) toolbox. All experiments in this section are run on an NVIDIA TESLA M40 GPU with 24 GB RAM, and the FPS is averaged over 100 runs.

Table 3 compares the performance of several SOTA models for object detection in terms of mean average precision (mAP), frames per second (FPS), floating point operations (FLOPs), and number of parameters. We first compare our model with Faster R-CNN with ResNet-50 backbone, which is a commonly used object detection model. Table 3 shows the performance comparison of the two models on the test set. Faster R-CNN has the highest mAP of 76.9% but has the lowest FPS of 4.5 and requires the highest number of FLOPs of 206.6 G and parameters of 41.1 M. SSD has a lower mAP of 59.9% but has a higher FPS of 10.8 and requires more FLOPs of 342.7 G and fewer parameters of 23.8 M compared to Faster R-CNN. RetinaNet has a slightly lower mAP of 68.0% and FPS of 4.9 than Faster R-CNN but requires similar FLOPs of 204.36 G and more parameters of 36.1 M. Cascade R-CNN has a higher mAP of 73.7% than RetinaNet but has the lowest FPS of 3.7 and requires more FLOPs of 234.46 G and parameters of 68.9 M. CornerNet has the lowest mAP of 57.8%, the lowest FPS of 0.4, and requires the highest number of FLOPs of 1765 G and parameters of 200.9 M among all the models. YOLOv5s has a mAP of 70.9%, a high FPS of 46.3, and requires relatively fewer FLOPs of 23.8 G and parameters of 9.1 M compared to all the other models except CornerNet. DCA-YOLOv8s, the proposed model, has a mAP of 72.9%, the highest FPS of 62.5, and requires a similar number of FLOPs of 24.9 G and parameters of 11.5 M as YOLOv5s but outperforms it in terms of mAP and is much faster in terms of FPS.

Overall, the proposed model DCA-YOLOv8s performs well in terms of both accuracy and speed compared to the other SOTA models for cattle detection.

4.4. Ablation study

In this section, we conduct an ablation study to analyze the contribution of each proposed improvement to the performance of our DCA-YOLOv8 model for cattle detection.

We first analyze the effect of deformable convolution on the performance of our model. Our initial investigation involves gradually replacing the convolutional layers in the C2f module of the YOLOv8 backbone with DCN to study its effects. Specifically, there are four C2f

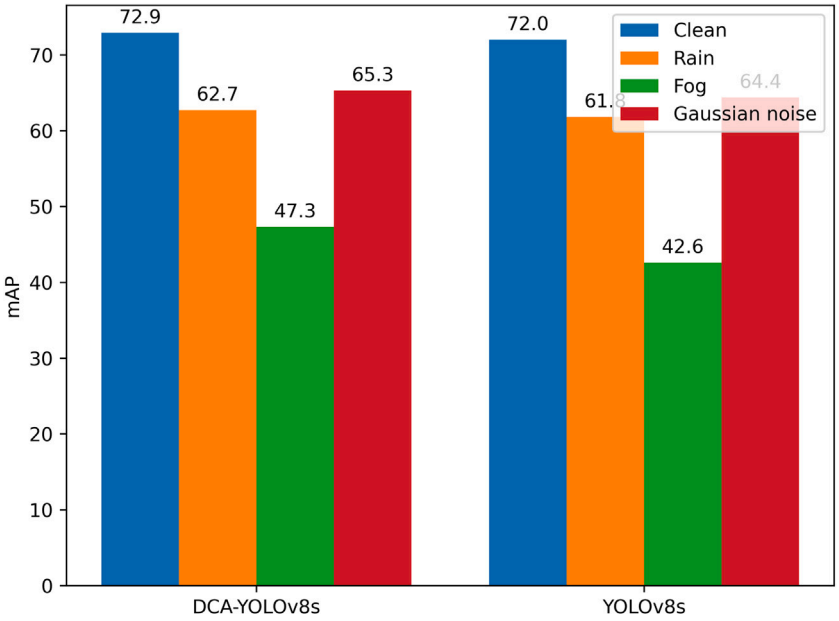

Fig. 11. Performance comparison of DCA-YOLOv8s and YOLOv8s under common corruptions.

Table 5
Comparison of different positions of CA module.

Small	Medium	Large	Model size (MB)	FLOPs (G)	mAP@0.5 (%)
✓			22.12	**24.9**	**72.9**
✓	✓		22.10	25.1	72.4
✓		✓	22.13	25.2	72.3
	✓	✓	22.14	25.2	72.1
✓	✓	✓	22.15	25.3	72.0

Table 6
Results of DCA-YOLOv8s with and without data augmentations used in this paper.

Data augmentations	mAP@0.5 (%)
✓	**72.9**
	72.5

modules in the backbone of YOLOv8, and we gradually replace the convolutional layers within them. Table 4 shows the performance comparison of the YOLOv8 model with and without deformable convolution on the test set. As can be seen, the inclusion of deformable convolution improves the mAP by 1.9%. These results demonstrate the effectiveness of adding deformable convolution to the C2f module in improving the performance of our model for cattle detection.

We also analyze the effect of coordinate attention on the performance of our model. Specifically, we added the CA module after the C2f module within the heads of different sizes and investigated its effects on each head. Table 5 shows the performance comparison of the YOLOv8s model with and without coordinate attention on the test set. As can be seen, including CA in the small head improves the mAP most. These results demonstrate the effectiveness of coordinate attention in improving the performance of our model for cattle detection.

Finally, we analyze the effect of data augmentation on the performance of our model. Table 6 shows the performance comparison of the DCA-YOLOv8s model trained with and without data augmentation on the test set. As can be seen, the inclusion of data augmentation improves the mAP by 0.4%. The results demonstrate the importance of data augmentation in improving the performance of our model for cattle detection.

Overall, these results demonstrate the effectiveness of each proposed improvement in improving the performance of our YOLOv8 model for cattle detection. In particular, deformable convolution and coordinate attention are effective techniques for improving the performance of object detection models, and data augmentation is an important technique for improving the robustness and generalization ability of deep learning models.

5. Discussion

5.1. Robustness under common corruptions

In this section, we conduct experiments to evaluate the robustness of our proposed method against image corruption. Image corruption can occur due to various factors such as environmental conditions, camera quality, and image processing techniques. It is important to evaluate the robustness of object detection models against these corruptions as it can impact their performance in real-world applications (Hendrycks and Dietterich; Li et al., 2022; Dong et al., 2023).

Due to the experimental farm's location in the southern region of China, where rain and fog weather conditions are likely to occur, we selected rain and fog corruptions for evaluation. Additionally, to evaluate the model's performance under various environmental noises, we tested the effect of Gaussian noise. We implement those corruptions following (Dong et al., 2023) with a severity level of 3.

Fig. 11 shows the mAP of the DCA-YOLOv8s and YOLOv8s under rain, fog and gaussian noise corruptions. As can be seen, the DCA-YOLOv8s improves the mAP on the corrupted images by 0.9%, 4.7%, and 0.9%, respectively. These results demonstrate the effectiveness of our proposed method in improving the robustness of the YOLOv8 model against common corruption in the agricultural environment.

Fig. 12 shows the detection results of the DCA-YOLOv8s model. As can be seen, the performance under fog weather degrades significantly. In addition, the model was able to detect cattle robustly in all other scenarios.

Overall, these experiments demonstrate the effectiveness of our proposed method in improving the robustness of the YOLOv8 model for cattle detection against image corruption. These results have important implications for real-world applications where the model needs to perform accurately in challenging conditions.

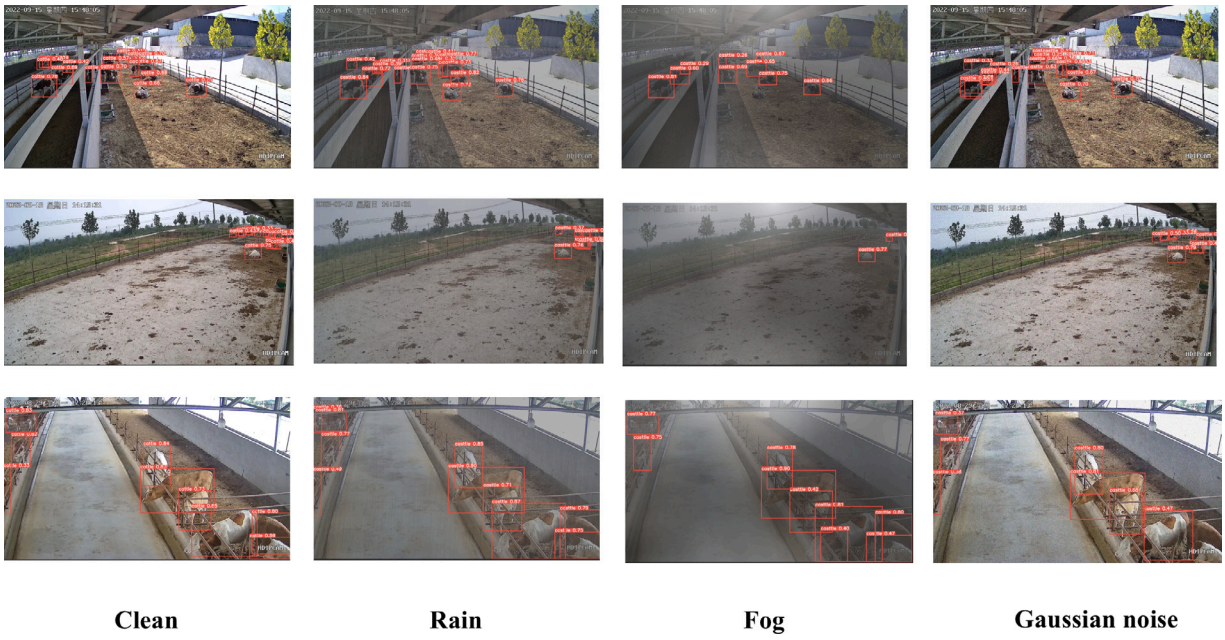

Fig. 12. DCA-YOLOv8s predictions under common corruptions.

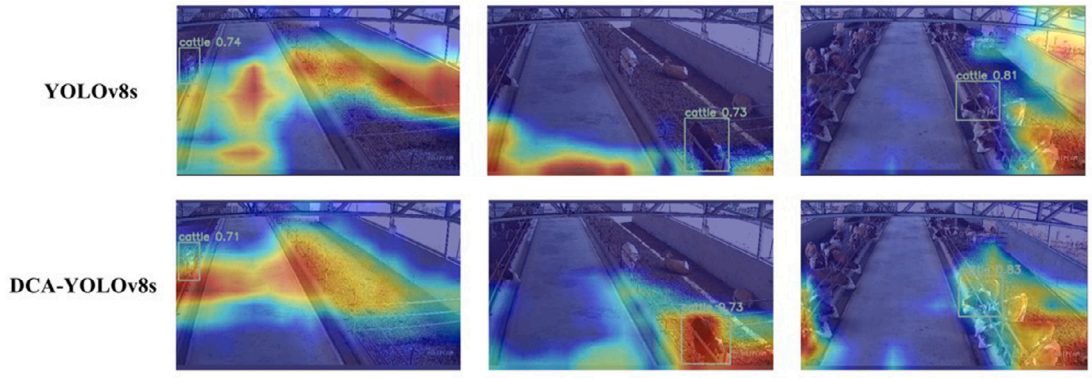

Fig. 13. Grad-CAM (Selvaraju et al., 2017) comparison of DCA-YOLOv8s and YOLOv8s.

5.2. Analysis of the detection results

In this section, we analyze the results of our proposed method for cattle detection using Grad-CAM (Selvaraju et al., 2017) visualization. Grad-CAM is a technique that visualizes the regions of an image that the model uses to make a prediction.

We use the YOLOv8s and DCA-YOLOv8s models trained on our cattle dataset and evaluate its performance on a subset of the test set. For each image in the subset, we generate a Grad-CAM visualization for the predicted bounding box of the cattle.

Fig. 13 shows some examples of the Grad-CAM visualizations generated by our proposed method. As can be seen, the model correctly identifies the cattle and generates accurate bounding boxes around them. The Grad-CAM visualizations show that the model focuses on the key features of the cattle, such as their heads, bodies, and legs. We also compare the Grad-CAM visualizations of our proposed method with those of the baseline YOLOv8s model without our proposed improvements. As can be seen in Fig. 13, the Grad-CAM visualizations of the baseline model are less precise and often include regions outside of the cattle. In contrast, the Grad-CAM visualizations of our proposed method are more focused and accurately capture the key features of the cattle. Moreover, it can be observed visually that the model does not give more attention to the distant cows, this may also explain a difficult problem in the field of object detection: objects that are farther away and smaller are more difficult to distinguish.

These results demonstrate the effectiveness of our proposed method in improving the accuracy and precision of the YOLOv8 model for cattle detection. The Grad-CAM visualizations provide insight into the regions of the image that the model uses to make its predictions and show that our proposed method is able to better focus on the key features of the cattle.

5.3. Limitations

One limitation of our proposed method is that it requires additional computation and memory resources due to the use of deformable convolution and attention mechanism. Future work can explore more efficient implementations of deformable convolution and attention mechanism to reduce the computational and memory costs. Moreover, our proposed method is currently only evaluated on a cattle dataset, and its performance on other object detection tasks needs to be investigated. In addition, future work can explore the combination of deformable convolution and attention mechanism with other object detection frameworks, such as Faster R-CNN (Ren et al., 2015)and Mask R-CNN (Ren et al., 2015), to further enhance the detection performance.

6. Conclusion

In this paper, we propose a novel approach for fast cattle detection using deformable convolution and coordinate attention within YOLOv8, and formed DCA-YOLOv8 detection model. Our proposed method enhances the YOLOv8 architecture by introducing deformable convolution to capture more fine-grained spatial information and coordinate attention to emphasize important features in the detection process. The experimental results demonstrate that our proposed method achieves a significantly higher mAP than the baseline model, which demonstrates the effectiveness of our approach in enhancing the detection accuracy for cattle detection. Future work can explore more efficient implementations and generalization of our proposed method to other object detection tasks.

CRediT authorship contribution statement

Wenjie Yang: Came up with the idea, Designed and conducted the experiments, Writing – original draft, Conception or design of the work. **Jiachun Wu:** Methodology, Analysis, Conception or design of the work. **Jinlai Zhang:** Project administration, Review, Analysis, Conception or design of the work. **Kai Gao:** Review, Analysis, Conception or design of the work. **Ronghua Du:** Review, Analysis, Conception or design of the work. **Zhuo Wu:** Review, Analysis, Conception or design of the work. **Eksan Firkat:** Review, Analysis, Conception or design of the work. **Dingwen Li:** Writing – review & editing, Analysis, Conception or design of the work.

Declaration of competing interest

The authors declare that there is no conflict of interest in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The project was supported by Innovation Project of Guangxi Graduate Education, China (YCBZ2021019) and the National Natural Science Foundation of China (grant number 61973047).

References

Banhazi, T.M., Lehr, H., Black, J., Crabtree, H., Schofield, P., Tscharke, M., Berckmans, D., 2012. Precision livestock farming: an international review of scientific and commercial aspects. Int. J. Agric. Biol. Eng. 5 (3), 1–9.

Cai, Z., Vasconcelos, N., 2019. Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 1. http://dx.doi.org/10.1109/tpami.2019.2956516, URL http://dx.doi.org/10.1109/tpami.2019.2956516.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. MMDetection: Open MMLab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155.

Chilukuri, D.M., Yi, S., Seong, Y., 2022. A robust object detection system with occlusion handling for mobile devices. Comput. Intell. 38 (4), 1338–1364.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 764–773.

Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J., 2023. Benchmarking robustness of 3D object detection to common corruptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1022–1032.

Fang, H.S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.L., Lu, C., 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Trans. Pattern Anal. Mach. Intell..

Girshick, R., 2015. Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1440–1448.

Glenn, J., 2022. YOLOv5 release v6.1. https://github.com/ultralytics/yolov5/releases/tag/v6.1.

Glenn, J., 2023. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hendrycks, D., Dietterich, T., Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations.

Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13713–13722.

Law, H., Deng, J., 2018. Cornernet: Detecting objects as paired keypoints. In: 15th European Conference on Computer Vision. ECCV 2018, Springer Verlag, pp. 765–781.

Li, S., Li, K., Qiao, Y., Zhang, L., 2022. A multi-scale cucumber disease detection method in natural scenes based on YOLOv5. Comput. Electron. Agric. 202, 107363.

Lin, T., 2015. LabelImg. Online: https://Github.Com/Tzutalin/LabelImg.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 21–37.

Qiao, Y., Guo, Y., He, D., 2023. Cattle body detection based on YOLOv5-ASFF for precision livestock farming. Comput. Electron. Agric. 204, 107579.

Qiao, Y., Guo, Y., Yu, K., He, D., 2022. C3D-ConvLSTM based cow behaviour classification using video data for precision livestock farming. Comput. Electron. Agric. 193, 106650.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 779–788.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. 28.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Su, D., Kong, H., Qiao, Y., Sukkarieh, S., 2021. Data augmentation for deep learning based semantic segmentation and crop-weed classification in agricultural robotics. Comput. Electron. Agric. 190, 106418.

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

Wang, Q., Cheng, M., Huang, S., Cai, Z., Zhang, J., Yuan, H., 2022b. A deep learning approach incorporating YOLO v5 and attention mechanisms for field real-time detection of the invasive weed Solanum rostratum Dunal seedlings. Comput. Electron. Agric. 199, 107194.

Wang, X., Zhao, Q., Jiang, P., Zheng, Y., Yuan, L., Yuan, P., 2022c. LDS-YOLO: A lightweight small object detection method for dead trees from shelter forest. Comput. Electron. Agric. 198, 107035.

Wei, H., Xu, E., Zhang, J., Meng, Y., Wei, J., Dong, Z., Li, Z., 2022. BushNet: Effective semantic segmentation of bush in large-scale point clouds. Comput. Electron. Agric. 193, 106653.

Weng, Z., Meng, F., Liu, S., Zhang, Y., Zheng, Z., Gong, C., 2022. Cattle face recognition based on a Two-Branch convolutional neural network. Comput. Electron. Agric. 196, 106871.

Zhang, J., Chen, L., Ouyang, B., Liu, B., Zhu, J., Chen, Y., Meng, Y., Wu, D., 2022. Pointcutmix: Regularization strategy for point cloud classification. Neurocomputing.

Zhang, J., Meng, Y., Wu, J., Qin, J., Yao, T., Yu, S., et al., 2020. Monitoring sugar crystallization with deep neural networks. J. Food Eng. 280, 109965.

Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J., 2023. Object detection in 20 years: A survey. Proc. IEEE.