Cow-YOLO: Automatic cow mounting detection based on non-local CSPDarknet53 and multiscale neck

De Li[1], Junhao Wang[1], Zhe Zhang[1], Baisheng Dai[1*], Kaixuan Zhao[2], Weizheng Shen[4*], Yanling Yin[1], Yang Li[3]

(1. *College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China*;
2. *College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang 471023, Henan, China*;
3. *College of Animal Sciences and Technology, Northeast Agriculture University, Harbin 150030, China*;
4. *Key Laboratory of Northeast Smart Agricultural Technology of Ministry of Agriculture and Rural Affairs, Northeast Agricultural University, Harbin 150030, China*)

Abstract: Cows mounting behavior is a significant manifestation of estrus in cows. The timely detection of cows mounting behavior can make cows conceive in time, thereby improving milk production of cows and economic benefits of the pasture. Existing methods of mounting behavior detection are difficult to achieve precise detection under occlusion and severe scale change environments and meet real-time requirements. Therefore, this study proposed a Cow-YOLO model to detect cows mounting behavior. To meet the needs of real-time performance, YOLOv5s model is used as the baseline model. In order to solve the problem of difficult detection of cows mounting behavior in an occluded environment, the CSPDarknet53 of YOLOv5s is replaced with Non-local CSPDarknet53, which enables the network to obtain global information and improves the model's ability to detect the mounting cows. Next, the neck of YOLOv5s is redesigned to Multiscale Neck, reinforcing the multi-scale feature fusion capability of model to solve difficulty detection under dramatic scale changes. Then, to further increase the detection accuracy, the Coordinate Attention Head is integrated into YOLOv5s. Finally, these improvements form a novel cow mounting detection model called Cow-YOLO and make Cow-YOLO more suitable for cows mounting behavior detection in occluded and drastic scale changes environments. Cow-YOLO achieved a precision of 99.7%, a recall of 99.5%, a mean average precision of 99.5%, and a detection speed of 156.3 f/s on the test set. Compared with existing detection methods of cows mounting behavior, Cow-YOLO achieved higher detection accuracy and faster detection speed in an occluded and drastic scale-change environment. Cow-YOLO can assist ranch breeders in achieving real-time monitoring of cows estrus, enhancing ranch economic efficiency.

Keywords: cows mounting, automatic detection, Cow-YOLO, computer vision, CSPDarknet53, multiscale neck
DOI: 10.25165/j.ijabe.20241703.8153

Citation: Li D, Wang J H, Zhang Z, Dai B S, Zhao K X, Shen W Z, et al. Cow-YOLO: Automatic cow mounting detection based on non-local CSPDarknet53 and multiscale neck. Int J Agric & Biol Eng, 2024; 17(3): 193–202.

1 Introduction

Timely detection of dairy cow estrus is the key to ensuring milk production and the economic benefits of cattle farms[1]. When cows are in estrus, cows will show characteristics such as increased activity, body temperature changes, and mounting behavior[2]. Cows mounting behavior is easy to observe and does not easily cause a stress response to the cows, so mounting behavior is the typical way to judge whether they are in estrus. With the development of sensor technology, many researchers have begun to use sensor technology to detect estrus behavior in dairy cows. Perez et al.[3] proposed using accelerometer ear tags to monitor both the temperature and activity level of cows, thereby enabling estrus detection in cows. Mičiaková et al.[4] proposed to use the Heatime RuminAct collar to monitor the activity level of cows for estrus detection. Wang et al.[5] proposed a combination of sensors and deep learning methods to detect the estrus behavior of dairy cows. However, the sensor acting as an attachment device can cause detachment and damage during the movement of the cow. It is even possible to cause a stress response to the cow, and it is also susceptible to interference from the behavior of other cows, leading to false detections. Traditional cows mounting behavior detection relies on manual observation, but this method cannot achieve real-time detection, which is easy to cause missed detection, resulting in low detection efficiency. Therefore, it is of great significance to study the automatic and real-time mounting behavior detection of dairy cows.

Some researchers proposed non-contact detection methods based on computer vision to discover mounting behavior of dairy cows. Tsai et al.[6] proposed a method to determine the cow's moving area of interest first, then use the foreground segmentation method to segment the moving cow, and finally judge the cows mounting behavior according to the length change of the moving

Received date: 2023-01-20 **Accepted date:** 2024-03-02
Biographies: De Li, MS, research interest: machine vision, Email: 1793446390@qq.com; **Junhao Wang**, MS, research interest: machine vision, Email: 1720036517@qq.com; **Zhe Zhang**, MS, research interest: machine vision, Email: 842372268@qq.com; **Kaixuan Zhao**, PhD, Professor, research interest: Smart animal husbandry, Email: kx.zhao@haust.edu.cn; **Yanling Yin**, PhD, Associate Professor, research interest: machine hearing, Email: yinyanling@neau.edu.cn; **Yang Li**, PhD, Associate Professor, research interest: ruminant nutrition, Email: liyang1405053@neau.edu.cn.

***Corresponding author: Baisheng Dai**, PhD Associate Professor. Research interest: intelligent animal husbandry. College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China. Tel: +86-13936253144, Email: bsdai@neau.edu.cn; **Weizheng Shen**, PhD, Professor, research interest: intelligent animal husbandry. College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China. Tel: +86-451-55191358, Email: wzshen@neau.edu.cn.

cow. Wang et al.[7] proposed a method that first uses an improved Gaussian mixture model to detect moving cows, then removes the interference background based on color and texture information, and at last, AlexNet[8] network was trained to recognize cows mounting behavior. Lodkaew et al.[9] proposed an automated estrus detection system (CowXNet) that relies on detecting mounting behaviors of cows to achieve an 83% accuracy rate in estrus detection. However, traditional computer vision technology has the problems of slow detection speed and low detection accuracy. Therefore, some researchers have proposed the use of object detection methods to detect the mounting behavior of cows. Chae et al.[10] proposed an improved YOLOV3[11] target detection method to detect cows mounting behavior. By introducing an additional layer and Mish activation function to YOLOV3, YOLOV3 can improve the performance of cows mounting behavior detection. Wang et al.[12] introduced DenseBlock, proposed a new boundary function and optimized anchor boxes to improve YOLOV3, thereby improving the ability of the YOLOV3 model to detect cows mounting behavior. However, several issues have not been addressed, since it is difficult to obtain the cows mounting feature in an occluded environment, which leads to a sharp drop in detection accuracy, moreover, since the cows are active in the pasture, the distance from the surveillance camera is not fixed, which will cause the scale to change drastically. These issues above make the model must have multi-scale detection ability. In addition, the detection speed of the proposed method is still unable to meet the real-time requirements.

In order to solve these problems, this paper proposes an efficient Cow-YOLO model to detect cows mounting behavior. Firstly, aiming at the problem that the accuracy of model detection decreases in occlusion environments, Non-local CSPDarknet53 which integrates GCNet[13] module and Swin Transformer[14] module is proposed to obtain global information to improve the recognition ability of models in occlusion environments. Secondly, due to the dramatic changes in the feature scale of dairy cows in dairy farms, the Neck module of YOLOV5[15] is unable efficiently perform multi-scale feature fusion, and the original neck module is redesigned to propose a Multiscale Neck. Finally, in order to further improve the accuracy of Cow-YOLO, Coordinate Attention Head which combines Coordinate Attention[16] is designed to accurately achieve the coordinate positioning of mounting cows. Cow-YOLO also inherits the advantages of fast recognition speed of YOLOV5s. These improvements make Cow-YOLO model is good at mounting behavior detection in occluded and drastic scale changes environments, and also allow the model to meet the real-time requirements.

The main contributions of this paper are: 1) To propose a novel cow mounting detection model named Cow-YOLO, achieving improved detection accuracy in environments with significant occlusion and scale variation, while meeting real-time requirements; 2) To integrate the GCNet module and the Swin Transformers module into the backbone to improve the detection ability of the model in an occluded environment; 3) To introduce the SPPF module and the BiFPN fusion method in the neck module, which can fuse features more efficiently to solve the problem of low detection rate at drastic scale changes environments; 4) To integrate the Coordinate Attention Heads into Cow-YOLO, which can further improve detection accuracy; 5) To construct a daily cow mounting dataset, and our Cow-YOLO surpasses most of the current mainstream methods on this dataset.

2 Datasets

2.1 Dataset construction of cows mounting

In order to improve the robustness of the model in the actual production environment, this paper no longer used a certain experimental ranch to collect the data of cows mounting behavior. The research data used in this study were data from videos and pictures of cows mounting behavior across the actual production environment on the Internet at https://github.com/IPCLabNEAU/C-ows-Mounting-Behavior-Detection. When collecting the cows mounting behavior data, this work collected as much as possible of the occluded cows mounting behavior data and scale changes drastically cows mounting behavior data to better solve the problem of low detection accuracy of the model under occlusion and dramatic scale changes. After preprocessing the acquired video data, the video decomposing frame technology was adopted, taking 1 frame every 5 frames to obtain the video frame image, and also adjusting the size of the video image to normalize the video frame image. This paper first collected 6247 images of cows mounting behavior. In addition, in order to better evaluate the detection ability in the occluded environments, this paper collected an additional 236 images of cows mounting behavior in the occluded environments. This paper has collected a total of 6483 cows mounting images as a dataset for the model, containing 3889 images of cows mounting in the occluded environments and 2594 images of cows mounting in the non- occluded environments, and some cows mounting images are shown in Figure 1. In the dataset, there are 4583 images with a single mounting behavior, 1900 images with multiple mounting behavior.

a. Images of cows mounting behavior in a without occluded environment

b. Images of cows mounting behavior in an occluded environment

Figure 1 Part of datasets of cows mounting behavior images

2.2 Data annotation

The use of a target detection model needs to provide the real position information of the cows mounting behavior during the training and testing process. In this study, the labelimg image annotation tool (https://github.com/tzutalin/labelImg) was used to label the 6483 cows mounting behavior datasets obtained. Then, the annotation information was saved as an XML file with the same name as the image according to the PASCAL VOC[17] format.

2.3 Dataset partition

The data set was randomly divided into 4048 training samples, 1012 validation samples and 1187 test samples using the 6247 cows mounting behavior images collected for the first time. In the training set, 90 non-mount images were added to optimize the performance of the model, resulting in a training set consisting of 4138 images. The test set containing 1187 test examples was named Test. In Test, there were 1187 images, of which the occluded cows mounting behavior images accounted for 37%. In addition, to better test the performance of the model, this study additionally constructed a test set called Test-high that removed 411 non-occluded cows mounting images in Test, and added previously collected additional 236 unused images of occluded cows mounting. In Test-high, there were a total of 1012 test examples, and 63.9% of the images in the occlusion environment were cows mounting images. Finally, in order to test the model's ability to test the mounting behavior of cows in a fully occluded environment, Test-high's 365 non-occlusion cows mounting behavior images were eliminated to form a fully occluded dataset called Test-challenge.

2.4 Data augmentation

The significance of data augmentation is to enrich the dataset and improve the robustness of the model. The hue, saturation, and brightness of the images were adjusted for some of the constructed cows mounting behavior images; some traditional data augmentation methods (zoom, translation) were used to augment the training set. In addition, Mosaic is a new data augmentation method that randomly mixes 4 training images, therefore, 4 different contexts are mixed. This allows the detection of objects outside their normal context and significantly reduces the need for a large mini-batch size[18]. Mosaic data augmentation is also used in Cow-YOLO, which finally brings the training set to 10 120 images by using several of the above data augmentation methods.

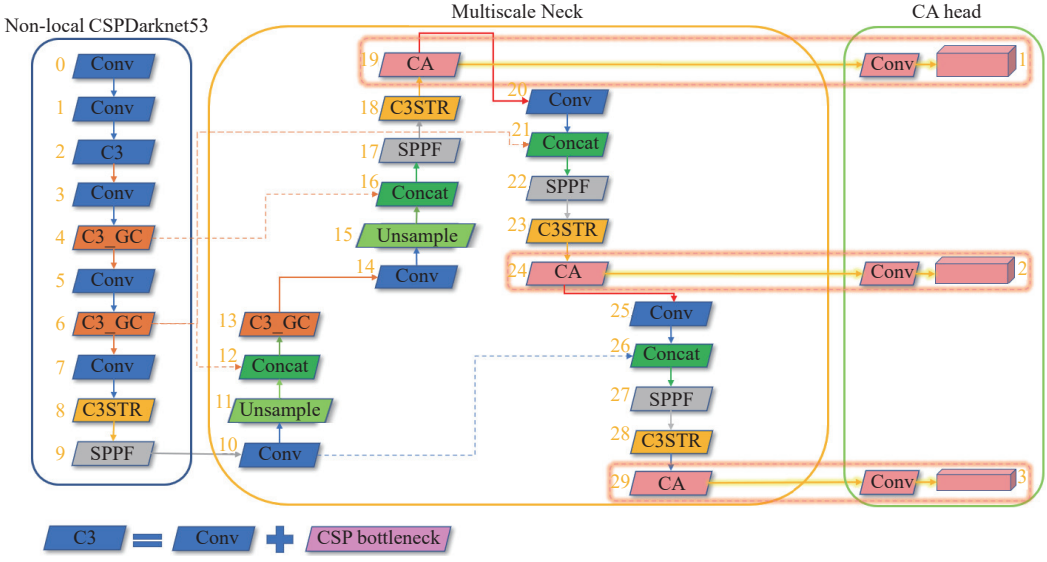
3 Methods

3.1 Overview of YOLOV5

The YOLOV5[15] model is a new single-stage object detection algorithm, which is an improved version of YOLOV4[18] with many advantages. YOLOV5 has 4 different baseline models, namely: YOLOV5s, YOLOV5m, YOLOV5l, and YOLOV5x. YOLOV5 uses CSPDarknet53 as the backbone, PANet[19] as the Neck, and the detection head used by YOLOV3[11]. Compared with previous generations of YOLO series algorithms[11,18,20,21], it can reduce the weight of the model while maintaining high accuracy. The detection speed of YOLOV5s exceeds 150 f/s, which is enough to meet the needs of real-time detection. Therefore, YOLOV5s is chosen as the baseline, from which Cow-YOLO was proposed.

3.2 Cow-YOLO

The overall structure of Cow-YOLO is shown in Figure 2. The original YOLOV5s[15] is improved to make it more suitable for cows mounting behavior detection in occluded and multi-scale environments, and meet real-time requirements. Firstly, in the backbone, Non-local CSPDarknet53 is proposed to obtain global information by introducing GCNet[13] and Swin Transformer[14] to the original CSPDarknet53. Secondly, in the Neck, Multiscale Neck is proposed to enhance the multi-scale feature fusion capability of Neck by adding a new multi-scale feature fusion method BiFPN[22], and adding the SPPF module. Finally, in the Head, Coordinate Attention Head is proposed to further improve the accuracy of model.



Note: Non-local CSPDarknet53 backbone contains Global Context Network blocks and Swin Transformer blocks. The Multiscale Neck integrates the structure of BiFPN and adds SPPF module. Coordinate Attention Head (CA Head) uses the Coordinate Attention blocks in the Head.

Figure 2 The overall structure of Cow-YOLO

3.2.1 Non-local CSPDarknet53

In object detection networks, many CNN backbones have been proposed[23-27], but CNN's backbones the inability to obtain global information. In the occlusion environment, there will be a lack of features, and the backbone based on CNN will aggravate this lack, which is very unfavorable for us to detect cows mounting behavior in the occlusion environment. Therefore, Non-local CSPDarknet53 was designed in order to solve the problem of difficult detection in occluded environments.

1) Global Context Network (GCNet) Block

GCNet is an attention mechanism with global context modeling ability. It inherits the advantages of the previous attention mechanism[28-30], which can be plug-and-play, lightweight, and model long-range dependencies. The experiments in Reference [13] have proved that adding GCNet Block can also improve the performance of the network and only add a few parameters. The overall structure is composed of two sub-modules, Context Modeling, and Transform, as shown in Figure 3. The feature map

first passes through a convolution layer and then a Softmax layer, and then performs matrix multiplication with the input feature map. Then the feature map enters the second sub-module, passes through a convolution layer again, then passes through LayerNorm and ReLU, and then performs a matrix addition operation with the original input feature map after convolution.

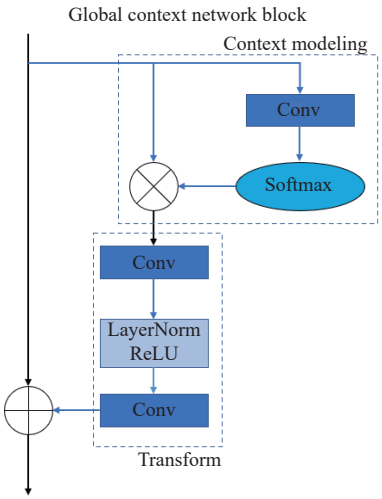


Figure 3 Structure of GCNet Block

Due to the fact that there are fewer cows mounting features that can be extracted in the occlusion environment, the effect of network learning will deteriorate. Therefore, GCNet Block is decided to replace the CSP bottleneck block, and GCNet Block is named C3_GC module in backbone. In our pre-experiment, adding the attention mechanism module at the beginning of the network did not improve the performance of the network, and even reduced the performance of the benchmark model, so GCNet Block is used in the middle of backbone to replace the CSP bottleneck block, as shown in Figure 2.

2) Swin Transformer Block

Inspired by transformer[15,31] are used in object detection task. The CSP bottleneck block in the original YOLOV5 was replaced by the Swin Transformer Block. Because Swin Transformer Block can obtain information on the global and different feature layers, which is conducive to extracting the characteristics of cows mounting behavior in the occlusion environment. Each Swin Transformer Block is composed of the structure shown in Figure 4. After the feature map is input, it will first go through the LayerNorm layer of the first module, and then go through a LayerNorm after a conventional multi-head self-attention mechanism and then it will go through a fully connected layer called MLP to the second module through the residual connection. Like the first module, the feature map will first go through a LayerNorm layer, and then through the multi-head self-attention mechanism with the ability to shift the window, then through the LayerNorm to the fully connected layer, and finally connect the next repeating module through the residual again. The Swin Transformer blocks in the network are named the C3STR module.

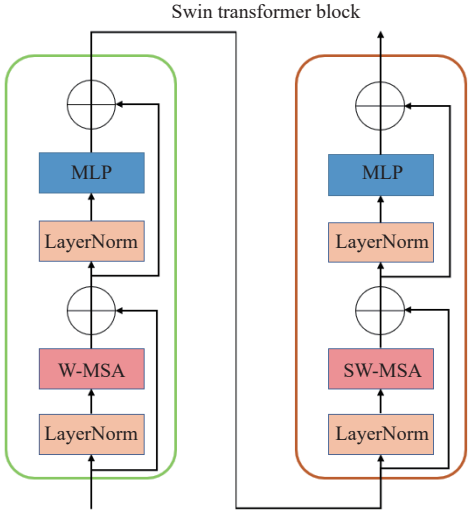


Figure 4 Structure of Swin Transformer Block

In Cow-YOLO, Swin Transformer Block is applied at the backbone end and the Neck end, as shown in Figure 2. Because the application at the end will reduce the training cost and will not bring too many parameters to the model so that the size of the model is controlled at a low level.

3.2.2 Multiscale Neck

In original YOLOV5 Neck, PANet structure with higher accuracy than FPN[32] was used for feature fusion, but PANet[19] also is difficult to adapt to the situation of drastic scale changes. In order to solve the problem that features are difficult to fuse under severe scale changes, Multiscale Neck is designed. Multiscale Neck adopts the feature fusion method of BiFPN and adds additional SPPF module.

1) Bi-directional Feature Pyramid Network (BiFPN)

BiFPN is a new multi-scale feature fusion method, which can make Cow-YOLO more suitable for detecting cows mounting behavior in environments with drastic scale changes. It takes into account the role of each node in the feature layer. There is only one input node in the feature layer. Its contribution to the feature fusion of the entire network is very small, and its removal will improve the efficiency of the entire feature layer fusion. In the feature layer of the same level, it adds an additional edge to the original input to the output node to strengthen the feature fusion capability of the same layer, so that the entire network can fuse more features and improve network performance.

The structure of BiFPN, as illustrated in Figure 5, eliminates nodes that have only one input edge, utilizing the top-down and bottom-up paths as feature network layers. The effectiveness of this structure has been proved in the experiments of Tan et al.[22], so this structure is adopted in the structure of Cow-YOLO to improve the multi-scale feature fusion ability of the model.

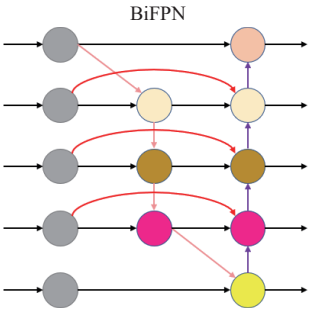


Figure 5 Structure of BiFPN

2) Spatial Pyramid Pooling Fast (SPPF)

Spatial Pyramid Pooling Fast not only maintains spatial information, but also converts feature inputs of different scales to the same scale, enabling the network to obtain multi-scale features without losing any information, and the execution speed is also faster than SPP[33]. Spatial information can ensure the recognition ability of Cow-YOLO in occluded environments, and the multi-scale information brought by SPPF enables Cow-YOLO to perform

well in the face of dramatic scale changes. Therefore, SPPF is added to the Multiscale Neck, which improves the detection ability of Cow-YOLO in occluded environments and in the case of drastic scale changes. The SPPF structure is shown in Figure 6.

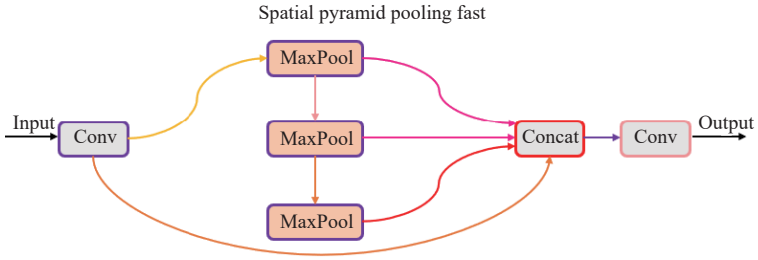


Figure 6 Structure of Spatial Pyramid Pooling Fast

After the feature map enters the SPPF block, it will first go through convolution and then enter three MaxPool layers with sizes of 5×5, 9×9, and 13×13 in turn. Whenever a new feature map is obtained through a MaxPool layer, in addition to continuing to input to the next MaxPool layer, it will also be directly output to Concat, and then perform the Concat operation together with the feature map that has not gone through MaxPool. Finally, the feature map output after passing Conv again, the position added in the Neck is shown in Figure 2.

3.2.3 Coordinate Attention Head

The original intention of the convolution head used by YOLOV5s is not to use it in an environment where only a small number of features can be obtained, which will make the model unable to make accurate detection in an occluded environment. Inspired by TPH-YOLOV5[34], the Coordinate Attention Head is proposed based on Coordinate Attention. Coordinate Attention can not only obtain global information and position information but also accurately highlight the area of interest, which greatly improves the network detection performance. The structure of Coordinate Attention is shown in Figure 7. The input feature map first passes through a residual block, and then obtains the horizontal and vertical position information through average pooling. The two feature maps containing location information will be turned into one feature map by Concat and Conv and then divided into two feature maps by BatchNorm and Non-linear. The two feature maps are again generated by Conv and Sigmoid to achieve coordinate attention generation. The last two feature maps and the feature map that only passes through the residual block are output after Re-weight operation.

The attention mechanism embeds position information into channel attention, which can not only obtain long-range dependencies along the spatial direction but also retain accurate position information, and finally generate a pair of direction-aware and position-sensitive attention maps. This mechanism makes it possible for the model to accurately locate the mounting cows, the Coordinate Attention Head is proposed as shown in Figure 2.

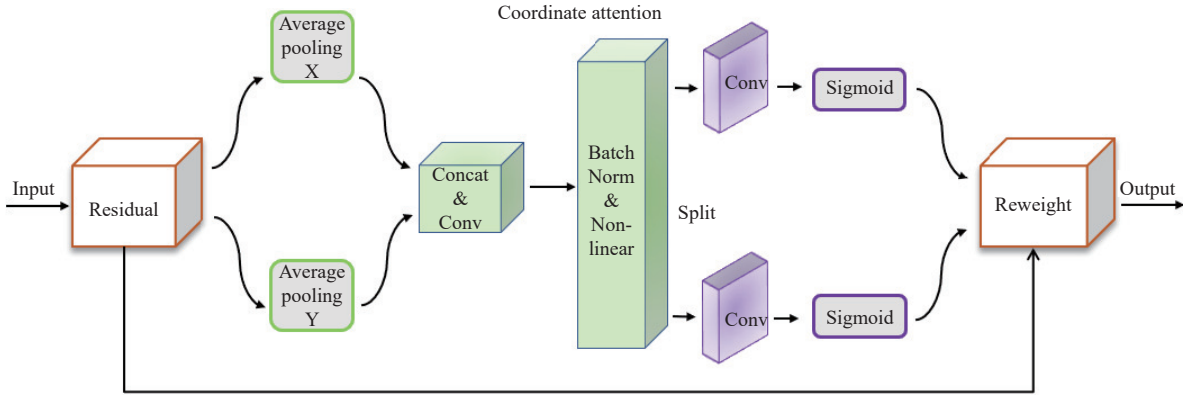


Figure 7 Structure of Coordinate Attention

3.3 Weighted box fusion

In object detection tasks, models typically return the location, class, and confidence score of detected objects. These three kinds of information are generally reflected by the given bounding box. There are generally three methods for model selection of bounding box[35,36]. But these methods cannot combine predictions from different models, which can improve detection accuracy[37]. Weighted Boxes Fusion (WBF) uses confidence scores of all proposed bounding boxes to construct average boxes, which enables WBF to combine predictions from different models. Therefore, this paper uses WBF to integrate the best model to further improve the detection accuracy of Cow-YOLO.

4 Experiments

4.1 Implementation details

This paper implemented Cow-YOLO in pytorch1.10.0 and all experimental models were trained and tested on NVIDIA GTX1080Ti GPU. During the model training process, due to GCNet Block in backbone could be well adapted to the pre-training weights obtained by YOLOV5 training on the COCO dataset[38], so the pre-training weights of YOLOV5s is utilized, which could save a lot of training time.

This paper used the pre-training weights to train 200 epochs on the training set, choosed the adam optimizer for training and the initial learning rate was set to 0.001. After training to the 102nd epochs, the early stop strategy was a technology that could avoid overfitting and ensured the best detection accuracy of the model, early stop strategy was adapted and continued training after adjusting the hyperparameters. The size of the image input by the model was 608×608. Because the size of the input image was not large, the batch size was set to 16. When training neural network models, the utilization of data augmentation was a common method

that could improve performance, so data augmentation was applied during training.

4.2 Evaluation metrics

To verify the effectiveness of the model proposed in this study, the following three metrics ar 21e used to evaluate the model: Precision, Recall and mean Average Precision (mAP). Precision is a measure of the accuracy of the target detection model's detection results for a certain category, that is, the proportion of the number of mounting behaviors detected by the model to the number of all detected targets. The calculation equation is:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \times 100\% \quad (1)$$

where, TP is the number of cows mounting behaviors that are correctly identified in the image, and FP is the number of non-mounting behaviors identified as mounting behaviors in the image.

Recall is a measure of the ability of the target detection model to find all the detection targets, that is, the ratio of the number of cows' mounting behaviors correctly identified by the model to the number of all mounting behaviors. The calculation equation is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (2)$$

where, FN is the number of cow mounting behaviors identified as non-mounting behaviors in the image.

mAP is also known as the mean Average Precision, the precision rate is the area enclosed by the *PR* curve and the coordinate axis, and the average precision rate is the average of the multi-category APs, that is, the integral of *P* to *R*. The calculation equation is:

$$\text{mAP} = \int_0^1 P(R)\,\mathrm{d}R \quad (3)$$

4.3 Experimental results and analysis

4.3.1 Compare with YOLO series of algorithms

In order to choose the best baseline in YOLO series of algorithms, YOLOV3-tiny, YOLOV3, YOLOV3-SPP, YOLOV4-tiny, YOLOV4, and YOLOV5s were trained, and the results of each model in the test are listed in Table 1.

Table 1 Comparison of Cow-YOLO and YOLO series detection methods

Methods	Precision/%	Recall/%	mAP/%	Speed/fps	Size/MB
YOLOV3	94.6	90.9	95.9	40.0	117.0
YOLOV3-tiny	90.6	85.8	90.7	368.0	16.6
YOLOV3-SPP	96.4	89.4	94.5	73.0	119.0
YOLOV4	96.9	98.0	98.4	62.0	245.0
YOLOV4-tiny	95.9	96.7	97.3	**371.0**	23.1
YOLOV5s	96.7	88.2	96.7	156.3	**13.7**
Cow-YOLO	**99.7**	**99.5**	**99.5**	156.3	18.0

Note: The bold numbers are used to highlight that they represent the maximum value for each column in the table. (The same below.)

Three evaluation indicators were used as the criteria for selecting the baseline. At the same time, the model size and detection speed were also taken into account. YOLOV5s has the advantages of fast recognition rate, small model size, and good accuracy. YOLOV5s was selected as the baseline of Cow-YOLO. Although YOLOV3 and YOLOV4 were outstanding representatives of the YOLO series of algorithms, Cow-YOLO outperformed them. Compared with YOLOV4 and YOLOV3, which were widely used in the industry, the speed were 152.1%, 290.8%, the mAP was improved 1.1%, 3.8%. As a new generation of lightweight YOLO model, YOLOV4-tiny also performed very well in Test. However, the size of the model was large, reaching 23.1 MB, so it was not suitable as a baseline. Compared with YOLOV4-tiny, Cow-YOLO reduced the model size by 22.1%, and improved Precision, Recall and mAP by 4.0%, 2.9%, and 2.3%, respectively. The comprehensive performance of Cow-YOLO proposed on the basis of YOLOV5s surpassed these YOLO series algorithms.

4.3.2 Compare with mainstream object detection methods

The current mainstream target detection methods can be divided into two categories, one is a two-stage target detection method represented by the RCNN series[39,40,41], and the other one is one-stage object detection methods such as YOLO and SSD[42]. Since the one-stage object detection method is more convenient and faster than the two-stage method, two-stage methods not are trained. In order to better compare the performance of Cow-YOLO in mainstream object detection algorithms, Test with more occlusions was selected as the test set. Besides, in addition to the YOLO series of algorithms trained previously, models widely used in one-stage object detection were also trained, such as YOLOX[43], RetinaNet[44], SSD (MobileNetV2), Efficentdet[22]. The performance of each model on the Test test set is listed in Table 2.

Table 2 Performance of each model on Test-high

Methods	Precision/%	Recall/%	mAP/%	Speed/f·s^{-1}	Size/MB
YOLOV3	95.5	89.2	92.4	40.0	117.0
YOLOV4	95.5	97.1	97.5	62.0	245.0
YOLOV4-tiny	96.5	95.0	96.1	**371.0**	23.1
YOLOV5s	96.2	85.9	96.2	156.3	**13.7**
Cow-YOLO	**99.4**	**99.0**	**99.5**	156.3	18.0
Efficentdet	96.3	95.9	96.2	97.0	14.9
RetinaNet	94.0	95.5	95.8	53.0	140.0
YOLOX-s	92.0	93.7	96.1	102.0	34.2
SSD(MobileNetV2)	96.6	82.2	95.7	59.0	14.2

In comparison with mainstream object detection methods, the Cow-YOLO also achieves the best results. Compared with Efficentdet, Cow-YOLO has a 2.6% mAP lead and a 61.0% faster speed. Cow-YOLO far exceeds new generation YOLO algorithm YOLOX-s, leading by 8.0%, 5.7% and 3.5% under the three evaluation indicators, respectively. And it was found that with the increase in the number of cows mounting behavior images in the occlusion environment, the mAP of most models decreased to varying degrees. Because the one-stage target detection model lacks the ability to obtain global information, the model cannot well extract cows mounting behavior features in the occluded environment. Therefore, it also proves that the realization of the Non-local mechanism is of great significance for the detection of cows mounting behavior in the occlusion environment. The scale changes too much during feature fusion, which makes the network miss some features. This is also the reason why the mainstream object detection methods miss the detection of small-scale cows mounting images. However, Cow-YOLO can perform good features in this complex multi-scale environment, and Multiscale Neck is indispensable.

4.3.3 Compare with the existing cows mounting detection methods

Compared with existing methods, the performance of Cow-YOLO can be evaluated more objectively. Since the detection accuracy published by the existing methods is the result obtained in almost no occlusion environment, the Cow-YOLO Test-high result is used as the comparison result. Table 3 lists the comparison between Cow-YOLO and existing cows mounting detection models.

Table 3 Comparison of Cow-YOLO and existing detection methods

Methods	Test sets (pieces)	Precision/%	Recall/%	mAP/%	Speed/f·s^{-1}
Cow-YOLO	1012 (occlusion rate 63.9%)	**99.40**	**99.00**	**99.50**	**156.30**
Wang et al.[12]	1440 (occlusion rate 1.9%)	99.15	97.62	--	31.00
LIU et al.[45]	5000 (occlusion rate 56%)	98.25	94.20	--	3.90
Guo et al.[46]	949	90.90	95.80	--	6.90
Noe et al.[47]	--	97.00	99.00	--	--

The test set of Wang et al.[12] also taked a video frame image every five frames, with a total of 1440 images, and the occlusion ratio is 1.9%. The test set of Liu et al.[45] was to directly decompose video frames, with 5000 mounting behavior images, and the occlusion ratio was 50%. The test set of Guo et al.[46] also directly decomposed video frames, with 949 mounting behavior images. Therefore, the result of Cow-YOLO in Test-high was chosen which occlusion ratio of 63.9% to compare with the above methods. Although directly decomposing video frames can obtain more images of the test set, it will reduce the diversity of test examples. When constructing a test set, decomposing video frames directly is not the best choice. Since most of the cows in the production environment do not show mounting behavior, which has brought enough negative examples to the test set, adding non-mounting images as negative examples will lead to inaccuracy in evaluating the performance of the model.

The traditional computer vision method adopted by Guo et al.[46] had high accuracy but a slow detection speed of only 6.9 f/s and the detection speed of Cow-YOLO was 22.7 times that of it. Wang et al.[12] used YOLOV3 as a baseline. Due to the bloated Darknet 53 network, this undoubtedly reducesd the detection speed. The detection speed was only 31 f/s, which also made the model not lightweight enough that the model could not be practically applied in dairy farms. The detection speed of Cow-YOLO led the method by 404.2%, and the model size was also much smaller than the method. Noe et al.[47] proposed a machine learning-based approach for detecting mounting behavior in cows, which also shows promising detection performance; however, this method is limited by the inherent constraints of machine learning. Cow-YOLO is proposed with many issues in mind, which is why Cow-YOLO can surpass the above methods.

4.3.4 Comparison of anti-occlusion capabilities

In order to better evaluate the recognition ability of the model in the occlusion environment, Test-challenge was chosen as the test set of the model's anti-occlusion ability. The methods that performed better before were selected to join the test, and the performance of each method under the Test-challenge test set are listed in Table 4.

Table 4 Performance of each model on Test-challenge

Methods	Test sets (pieces)	Precision/%	Recall/%	mAP/%	Speed/f·s^{-1}
YOLOV4	647	93.60	95.40	95.30	62.0
YOLOV5s	647	95.70	83.00	94.80	**156.3**
Cow-YOLO	647	**99.10**	**99.40**	**99.50**	**156.3**
Efficentdet	647	95.20	94.70	95.70	74.0
Wang et al.[12]	120	90.23	--	--	31.0
Liu et al.[45]	2817	92.15	--	--	3.9

Liu et al.[45] obtained it on the test set of 2817 occluded cows mounting images, but the data set directly decomposed the video frames without taking interval sampling. Wang et al.[11] and other methods in Table 4 took one frame of occluded cows mounting images at intervals of five frames as the test image. The test set of Wang et al.[12] has 120 images, and the other methods in Table 4 have 647 images.

Although their model still had good performance under a small amount of occlusion, in a fully occluded environment, the method precision of Wang et al.[12] decreased by 9.9%, and the method of Liu et al.[45] decreased by 6.6%. The mAP of YOLOV4 also dropped from 98.4 to 95.3, with a drop of 3.3%. As the baseline YOLOV5s, its recall dropped by 6.3%, and its mAP dropped by 2.0%. One of the essential reasons for the performance degradation is that the backbone for obtaining global information is not implemented. Cow-YOLO has considered the occlusion problem at the beginning of its design, so Cow-YOLO still maintains good performance. The iteration diagram of Precision, Recall and mAP during the training process is shown in Figure 8.

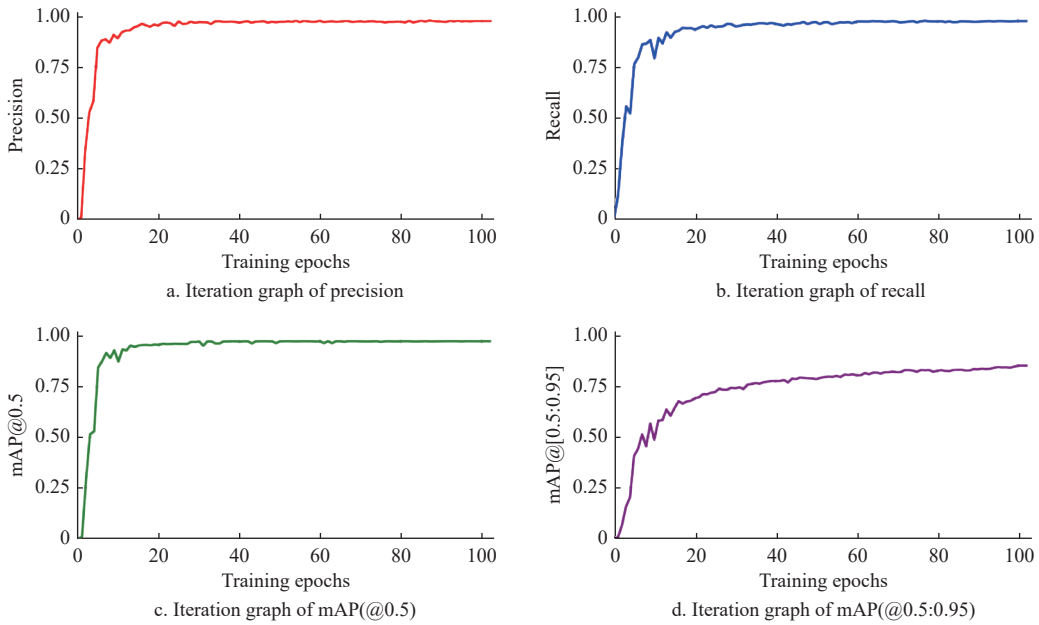
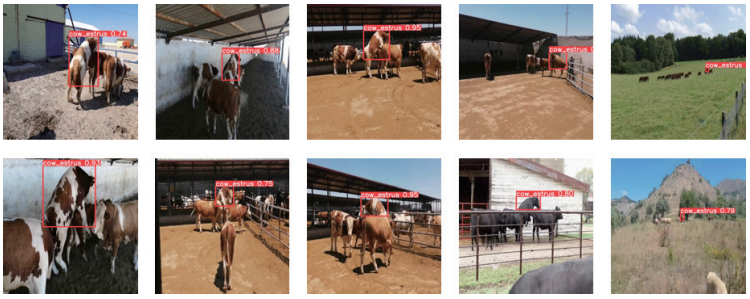


Figure 8 Iteration graph of Precision, Recall and mAP

Since the multi-scale problem is a common problem when collecting datasets, this paper does not conduct experiments on multi-scale problems separately. The multi-scale detection results are shown in Figure 9 and the detection results under complex light environment are shown in Figure 10.

In Figure 11, this study shows some of the detection results of other comparative detection methods in severe occlusion environments and medium and long distances.



a. a. Detection results of cowsmounting behavior at closedistances.

b. b. and c. Detection results of cows mounting behavior at medium distances.

c.

d. d. Detection results of cows mounting behavior at far distances.

e. e. Detection results of cows mounting behavior at ultra-long distances

Figure 9 Detection results of Cow-YOLO in the environment of severe scale transformation

a. a. and b. Are the detection result in the occluded environment.

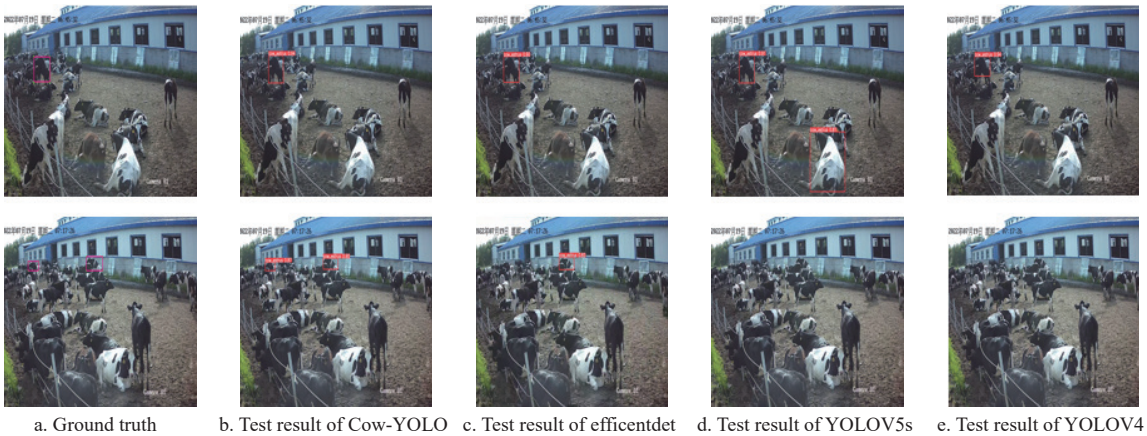
b.

c. c. Is the image taken with the phone in ranch.

d. d. Is the image taken by the ranch surveillance during the day.

e. e. Is the image taken at night by ranch surveillance

Figure 10 Detection results of Cow-YOLO in complex lighting environment



a. Ground truth b. Test result of Cow-YOLO c. Test result of efficentdet d. Test result of YOLOV5s e. Test result of YOLOV4

Figure 11 Detection results of partly comparative detection methods

As can be seen from Figure 11c-11e, in severe occlusion environments, the model that previously performed well in the test-challenge was unable to successfully detect. The most important reason is the lack of ability to obtain global information. Moreover, Figures 11d and 11 did not accurately mark the mounting cows in the non-occlusion environment because of the lack of improvement in adapting to the environment with severe scale changes

Cow-YOLO not only performs well in occlusion environments but also has strong adaptability under drastically multi-scale transformations. Furthermore, due to the ability of Non-local CSPDarknet-53 to obtain global information, the excellent feature fusion ability of Multiscale Neck and the accurate positioning of Coordinate Attention Head, Cow-YOLO is still accurate to detect cows mounting behavior in complex lighting conditions.

4.3.5 Ablation studies

Under Test-high, the importance of each improvement component was analyzed. The impact of each component on model performance is listed in Table 5.

Table 5 Ablation experiments under the Test-high dataset

Methods	mAP/%	GFLOPs
YOLOV5s	96.2	15.9
YOLOV5s+BiFPN	96.4(↑0.2)	16.1
YOLOV5s (previous + GCNet)	96.8(↑0.4)	17.2
YOLOV5s (previous + SPPF)	97.1(↑0.3)	19.8
YOLOV5s (previous+ Swin Transformer)	98.8(↑1.7)	110.9
Cow-YOLO (previous + Coordinate Attention)	99.5(↑0.7)	110.9

1) Improved feature fusion method

After YOLOV5 feature fusion method was changed PANet to BiFPN, this increased GFLOPs from 15.9 to 16.1, but mAP increased from 96.2 to 96.4 and enabled Neck to obtain a new feature path from backbone. This feature fusion method allows the Neck part of Cow-YOLO to obtain more features than the Neck part of YOLOV5s, which is beneficial to other components.

2) Effects of Swin Transformer Block

After Swin Transformer Block was added, the GFLOPs increased from 19.8 to 110.9 due to the larger computation required by the block, and the training overhead was also increased, but the mAP increased by 1.7. Among all the modules, this module contributes the most to the performance improvement of the model. In order to better recognize the recognition ability in the occluded environment, it is worthwhile to increase the amount of calculation.

3) Effects of Coordinate Attention Head

The Coordinate Attention Head proposed has almost no change to the GFLOPs of the model. The original intention of the Coordinate Attention mechanism is also to apply it in mobile networks. It can make the detection head pay more attention to the mounting behavior of cows, and improve the upper limit of the ability of Cow-YOLO to recognize the mounting of cows.

5 Conclusions

In this work, in order to solve the detection of cows mounting behavior in the environment of occlusion and severe scale changes, and meet the needs of real-time detection, this study proposed a new cows mounting detection model called Cow-YOLO. The proposal of Non-local CSPDarknet53 enables Cow-YOLO to obtain global information, which solves the problem of difficult detection of cows mounting behavior in occluded environments; the ingenious design of Multiscale Neck also solves the problem of difficult detection of mounting behavior under severe scale changes; the use of Coordinate Attention Head further achieves higher the detection accuracy. The experimental results show that Cow-YOLO can effectively detect the mounting behavior of cows in the case of occlusion and severe scale transformation, with high accuracy and high speed in detecting mounting behavior. On the test set, Cow-YOLO has a mean Average Precision of 99.5% and a detection speed of 156.3f/s, which still achieves better detection results compared to other methods, so this proves that the Cow-YOLO model effectiveness. The non-contact detection model of cows mounting proposed in this paper will not cause stress response to cows and is more beneficial to animal welfare. It provides a new option for cows mounting detection, improves the detection efficiency of cows mounting behavior, and further realizes the automation and intelligence of cows mounting detection. The future work needs to improve the ability of model to detect cows mounting behavior from videos, so that the model can be more conveniently applied in production.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32072788, 31902210, 32002227, 32172784), the National Key Research and Development Program of China (Grant No. 2019YFE0125600) and the earmarked fund (Grant No. CARS36).

[References]

[1] Higaki S, Okada H, Suzuki C, Sakurai R, Suda T, Yoshioka K. Estrus detection in tie-stall housed cows through supervised machine learning using a multimodal tail-attached device. Computers and Electronics in Agriculture, 2021; 191: 106513.

[2] Reith S, Hoy S. Behavioral signs of estrus and the potential of fully automated systems for detection of estrus in dairy cattle. Animal, 2018; 12(2): 398–407.

[3] Perez Marquez H J, Ambrose D J, Bench C J. Behavioral changes to detect estrus using ear-sensor accelerometer compared to in-line milk progesterone in a commercial dairy herd. Frontiers in Animal Science, 2023; 4: 1149085.

[4] Mičiaková M, Strapák P, Strapáková E. The influence of selected factors on changes in locomotion activity during estrus in dairy cows. Animals, 2024; 14(10): 1421.

[5] Wang J, Bell M, Liu X H, Liu G. Machine-learning techniques can enhance dairy cow estrus detection using location and acceleration data. Animals, 2020; 10(7): 1160.

[6] Tsai D M, Huang C Y. A motion and image analysis method for automatic detection of estrus and mating behavior in cattle. Computers and Electronics in Agriculture, 2014; 104: 25–31.

[7] Wang S H, He D J, Liu D. Automatic Recognition Method of Dairy Cow Estrus Behavior Based on Machine Vision. Transactions of the Chinese Society for Agricultural Machinery, 2020; 51(4): 241–249. (in Chinese)

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutionnal neural networks. Communications of the ACM, 2017; 60(6): 84–90.

[9] Lodkaew T, Pasupa K, Loo C K. CowXNet: An automated cow estrus detection system. Expert Systems with Applications, 2023; 211: 118550.

[10] Chae J-w, Cho H-c. Identifying the mating posture of cattle using deep learning-based object detection with networks of various settings. Journal of Electrical Engineering & Technology, 2021; 16: 1685–1692.

[11] Redmon J, Farhadi A. Yolov3: An incremental improvement. Computer Science, 2018; In press. doi: 10.48550/arXiv.1804.02767.

[12] Wang S H, He D J. Estrus behavior recognition of dairy cows based on improved YOLOv3 model. Transactions of the CSAE, 2021; 52(7): 141–150.

[13] Cao Y, Xu J R, Lin S, Wei F Y, Hu H. GCNet: Non-local networks meet squeeze-excitation networks and beyond. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South): IEEE, 2019; doi: 10.1109/ICCVW.2019.00246.

[14] Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada: IEEE, 2021; pp.9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[15] Jocher G. v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support. 2021; Available: https://github.com/ultralytics/yol-ov5/releases/tag/v6.0. Accessed on [2023-10-20].

[16] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for efficient mobile network design. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, 2021; pp.13708–13717. doi: 10.1109/CVPR46437.2021.01350.

[17] Everingham M, Van Gool L, Williams C K I, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. International Journal of Computer Vision, 2010; 88: 303–338.

[18] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection. Computer Science, 2020; doi: 10.48550/arXiv.2004.10934.

[19] Liu S, Qi L, Qin H F, Shi J P, Jia J Y. Path aggregation network for instance segmentation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, 2018; pp.8759–8768.

[20] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016; pp.779–788. doi: 10.1109/CVPR.2016.91.

[21] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE, 2017; pp.6517–6525. doi: 10.1109/CVPR.2017.690.

[22] Tan M X, Pang R M, Le Q V. Efficientdet: Scalable and efficient object detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, 2020; pp.10778–10787.

[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Computer Science, 2014; In press. doi: 10.48550/arXiv.1409.1556.

[24] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, 2016; pp.770–778.

[25] Howard A G, Zhu M L, Chen B, Kalenichenko D, Wang W J, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. Computer Science, 2017; In press. doi: 10.48550/arXiv.1704.04861.

[26] Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, 2018; pp.4510–4520. doi: 10.1109/CVPR.2018.00474.

[27] Tan M X, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. Inter-national conference on machine learning. PMLR, 2019; 97: 6105–6114.

[28] Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018; 42(8): 2011–2023.

[29] Woo S, Park J, Lee J-Y, Kweon I S. Cbam: Convolutional block attention module. In: Computer Vision - ECCV 2018, Springer, 2018; pp.3–19. doi: 10.1007/978-3-030-01234-2_1.

[30] Wang X L, Girshick R, Gupta A, He K M. Non-local neural networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA: IEEE, 2018; pp.7794–7803. doi: 10.1109/CVPR.2018.00813.

[31] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv: 2010.11929, 2020; In press. doi: 10.48550/arXiv.2010.11929.

[32] Lin T-Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA: IEEE, 2017; pp.936–944. doi: 10.1109/CVPR.2017.106.

[33] He K M, Zhang X Y, Ren S Q, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015; 37(9): 1904–1916.

[34] Zhu X K, Lyu S C, Wang X, Zhao Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada: IEEE, 2021; pp.2778–2788.

[35] Neubeck A, Van Gool L. Efficient non-maximum suppression. In: 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China: IEEE, 2006; pp.850–855.

[36] Bodla N, Singh B, Chellapp.R, Davis L S. Soft-NMS - improving object detection with one line of code. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy: IEEE, 2017; 5562–5570.

[37] Solovyev R, Wang W, Gabruseva T. Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing, 2021; 107: 104117.

[38] Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. Computer Vision - ECCV 2014, Springer, 2014; pp.740–755. doi: 10.1007/978-3-319-10602-1_48.

[39] Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE, 2015; pp.1440–1448. doi: 10.1109/ICCV.2015.169.

[40] Ren S Q, He K M, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 39(6): 1137–1149.

[41] He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017; pp.2980–2988. doi: 10.1109/ICCV.2017.322.

[42] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C-Y, et al. Ssd: Single shot multibox detector. In: Computer Vision - ECCV 2016, Springer, 2016; pp.21–37. doi: 10.1007/978-3-319-46448-0_2.

[43] Ge Z, Liu S T, Wang F, Li Z M, Sun J. Yolox: Exceeding yolo series in 2021. arXiv: 2107.08430, 2021; In press. doi:10.48550/arXiv.2107.08430

[44] Lin T-Y, Goyal P, Girshick R, He K M, Dollár P. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 42(2): 318–327.

[45] Liu Z C, He D J. Recognition method of cow estrus behavior based on convolutional neural network. Transactions of the Chinese Society for Agricultural Machinery, 2019; 50(7): 186–193. (in Chinese)

[46] Guo Y Y, Zhang Z R, He D J, Niu J Y, Tan Y. Detection of cow mounting behavior using region geometry and optical flow characteristics. Computers and Electronics in Agriculture, 2019; 163: 104828.

[47] Noe S M, Zin T T, Tin P, Hama H. Detection of estrus in cattle by using image technology and machine learning methods. In: 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), Kobe, Japan: IEEE, 2020; pp.320–321. doi: 10.1109/GCCE50665.2020.9291987.