Computers and Electronics in Agriculture 204 (2023) 107579

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Cattle body detection based on YOLOv5-ASFF for precision livestock farming

Yongliang Qiao [a], Yangyang Guo [b,*], Dongjian He [c]

[a] Australian Centre for Field Robotics (ACFR), Faculty of Engineering, The University of Sydney, NSW 2006, Australia
[b] School of internet, Anhui University, Hefei, Anhui 230039, China
[c] College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China

ARTICLE INFO

Keywords:
Cattle
YOLOv5 model
Attention mechanism
Intelligent monitoring

ABSTRACT

Precision livestock farming is a hot topic in the field of agriculture at present. However, due to the diversity of breeding environments, the current intelligent monitoring of animal information still faces challenges. In this study, a YOLOv5-ASFF object detection model was proposed to detect cattle body parts (e.g. individual, head, legs) in complex scenes. The proposed YOLOv5-ASFF consists of two components: YOLOv5 responsible for extracting multi-scale features from sample images, while ASFF was used to adaptively learn fused spatial weights for each scale feature map and fully acquire the features. In this way, the cattle area detection was realized and the generalization of detection model was improved. To verify the applicability and robustness of YOLOv5-ASFF, a challenging dataset consisting of cattle (cow and beef) with complex environments (e.g. different lighting, occlusion, different depths of field, multiple targets and small targets) was constructed for experimental testing. The proposed method based on YOLOv5-ASFFachieved a precision of 96.2%, a recall of 92%, an F1 score of 94.1%, and an mAP@0.5 of 94.7% on this dataset, which outperformed Faster R-CNN, Cascade R-CNN, SSD, YOLOv3 and YOLOv5s. Experimental results showed that the YOLOv5-ASFF method could fully learn more animal biometric visual features, thereby improving the performance of cattle detection model, especially the detection of key parts. Overall, the proposed deep learning-based cattle detection method is favorable for long-term autonomous cattle monitoring and management in intelligent livestock farming.

1. Introduction

Precision agriculture plays an important role in providing food, meat and fiber for human survival. Due to population growth, urbanization and lack of resources, human demand for food is getting higher and higher. As the second largest food supplier, Livestock production needs to improve its feeding efficiency and precise management to improve the productivity of each animal and meet the growing global demand for livestock products (Qiao et al., 2021; Fournel et al., 2017). In the recent era, intelligent equipments and artificial intelligence technology have become one of the main frameworks for precision livestock farming, which focusing on minimizing environmental impact and simultaneously maximizing livestock products (Mishra et al., 2021; Lim et al., 2020).

In the development of precision livestock farming(PLF), the automatic, efficient and accurate acquisition of agricultural animal information has always been a key prerequisite. For example, livestock farming robots (Zhang et al., 2020), Internet of Things technology (Patil and Patil, 2020), remote sensing technology (Liu et al., 2021a) and other technologies are used in the agricultural and livestock information monitoring application. At present commonly used methods to obtain farm information are sensors (Gertz et al., 2020), images or videos (Guo et al., 2019, 2021; Qiao et al., 2022), or a combination of the two (Sun et al., 2021). Sensor-based animal information acquisition methods often need to install corresponding sensors on animals, which may lead to stress reaction of animals and affect animal health and welfare. In addition, sensor failure or physical movement of animal collars lead to sensor repositioning, resulting in data loss or deviation (Wang et al., 2021b; Peng et al., 2019). The analysis technology based on image or video, as a non-contact technology, can complete the automatic monitoring and perception of animal behavior without stress. More recently, deep learning with automatic feature extraction and powerful image representation capabilities enables vision-based object detection and behavior recognition to be achieved more efficiently, and thus is also widely used in PLF (Qiao et al., 2019b,a).

In addition, the perception of animal phenotype information is of great significance for breeding management and breed selection. Detection of animal and its key body parts (e.g. back, head, legs) is conducive to monitor animal welfare information, and the relative position of body key parts can further reflect animal pose and behavior situation (Beggs et al., 2019). For example, the head area can be

* Corresponding author.
E-mail address: guoyangyang113529@ahu.edu.cn (Y. Guo).

https://doi.org/10.1016/j.compag.2022.107579
Received 28 July 2022; Received in revised form 10 November 2022; Accepted 15 December 2022
Available online 20 December 2022

used to judge the movement direction; the legs area can be used to judge whether it is lame; the back area can be used to evaluate the body condition. Therefore, the combination of deep learning and video technology to detect key parts of animal is of great significance for autonomous management. However, it is still challenging to achieve accurate detection of key parts of animals in a complex farm environment (e.g. occlusion, illumination) (He et al., 2016). Riekert et al. (2020) used deep learning system to detect pig positions and postures from 2D camera images, and the average precision has reached more than 80.2%. Shao et al. (2020) proposed a cattle detection and counting system based on Convolutional Neural Networks (CNNs), obtained a precision of 0.957. Jiang et al. (2019) proposed FLYOLOv3 deep learning to detect key parts (e.g. trunk, leg and head) of individual dairy cow. Although the above approaches demonstrated the feasibility of deep learning-based animal detection, it is still challenging to achieve accurate detecting of key parts of cattle in a complex farm environment (e.g. multi-cows, occlusion, different illumination, day and night) (He et al., 2016).

More recently, YOLO (You Only Look Once) network, one-stage object detection algorithm, which only uses a single network to process images and directly calculate object position (Shafiee et al., 2017). Object recognition systems from the YOLO family are often used for recognition tasks (Overall target recognition, local area recognition), and have been shown to outperform other target recognition algorithms (Jiang et al., 2019; Yan et al., 2021). YOLOv5 has proven to significantly improve the processing time of deeper networks (Zhu et al., 2021; Zhao et al., 2021). However, since YOLO algorithms do not need to generate candidate boxes, the detection accuracy is lower than Two-Stage algorithm. In addition, there is a problem of misdetection or omission in occlusion or small target detection. The adaptively spatial feature fusion (ASFF) is a kind of adaptive fusion of spatial features between different scales to solve the impact of target scale changes in target detection and achieve multi-scale and small target detection (Feng and Yi, 2022; Liu et al., 2019). In this study, in order to improve remote animal detection accuracy, we explored the application of deep learning in animal detection using the cutting-edge object detection framework YOLOv5 and the adaptively spatial feature fusion (ASFF). Here we proposed YOLOv5-ASFF model to detect the cattle. Firstly, YOLOv5 was used to extract multi-scale features from sample images. Then, ASFF was used to adaptively learn fused spatial weights for each scale feature map and enhance the animal detection performance. In our detection experiments, the whole cattle (cow and beef), the head, and the legs were detected respectively.

The contributions of this paper are as follows: (1) We integrated the ASFF into YOLOv5, proposed YOLOv5-ASFF based approach for cattle body parts detection. the proposed YOLOv5-ASFF can fully learn the more animal bio-metric visual features and improve the performance of cattle detection, especially key parts detection. (2) A challenging dataset consisting of cattle (cow and beef) with complex environments (e.g. different lighting, occlusion, different depths of field, multiple targets and small targets) was constructed for experimental testing. Experimental results show that the proposed YOLOv5-ASFF based approach achieved a precision of 96.2%, a recall of 92%, an F1 score of 94.1%, and an mAP@0.5 of 94.7%, which outperformed Faster R-CNN, Cascade R-CNN, SSD, YOLOv3 and YOLOv5s. (3)The test running speed of YOLOv5-ASFF is 41 frame per second, which satisfy the real-time requirement of PLF applications. Overall, the proposed YOLOv5-ASFF provides a real-time and high accuracy cattle body parts detection approach in complex scenes, which facilitates long-term autonomous cattle monitoring and management in intelligent livestock farming.

2. Material and methods

2.1. Data acquisition

In our experiments, the cattle dataset consists of cow dataset and beef dataset. The cow data was acquired from a commercial dairy farm (Yangling Keyuan Cloning Co., Ltd. with about 300 cows) in Yangling, China. See Guo et al. (2021). For cow data, the camera was positioned on the supported beam of a feeding shed at a height of 1.8 m and 35 m away from the aisle to get a full day video of cows walking in the aisle. The beef dataset was extracted from videos that were captured using a smartphone (iPhone8 Plus) under field conditions at the Animal Husbandry Teaching Test Base, Northwest Agriculture and Forestry University, Yangling, China. See Li et al. (2019). For beef data, the smartphone position was not fixed to obtain the daytime dataset of beef with different angles and scales. Therefore, the data set of cattle is relatively complex. The data samples are shown in Fig. 1.

The dataset obtained is challenging: multi-cattle appeared and exist occlusion; The light changed from dawn to dusk; Complex backgrounds included crush, soil ground, building background, etc. A total of 1000 images (cow: 500, beef: 500) were acquired. Among these images, a randomly selected 700 images were used to train the network, while the remaining 300 images were regarded as the testing dataset. In our experiments, the whole cattle, head, and legs were labeled manually with bounding boxes for testing the detection of key body parts.

2.2. YOLOv5-ASFF model for detecting cattle

The overall structure of the proposed method is illustrated in Fig. 2. It is improved based on the detection network of YOLOv5s, and it can be divided into three parts: backbone, neck, and prediction (Wang and He, 2021). YOLOv5 adopts the Path Aggregation Network (PANet) (Liu et al., 2018) structure, which leads to insufficient fusion of multi-scale features. Therefore, we introduce Adaptively Spatial Feature Fusion (ASFF) (Liu et al., 2019) structure to make full use of the semantic information of the high-level features of the image and the finegrained features of the bottom layer, and fully integrates the features by learning the weight parameters to enhance the fusion effect.

YOLOv5-ASFF network is mainly composed of backbone part, neck network, and prediction part: (1) The backbone of YOLOv5 is responsible for extracting target features, which includes Focus, Conv, C3, and Spatial Pyramid Pooling (SPP) layer; (2) The neck is a combination structure of a feature pyramid network (FPN) (Lin et al., 2017) and a path aggregation network (PANet) (Liu et al., 2018), which transmits the low-level feature information to the high-level feature maps through bottom-up paths, and high-level semantic information is sent to the bottom feature maps through top-down approaches; (3) The prediction part is ASFF-head module to achieve full fusion of image multi-scale features by learning the fusion method of weight parameters.

2.2.1. Backbone of YOLOv5-ASFF model

Backbone network is a convolutional neural network which aggregates different fine-grained images and forms image features, and includes Focus, Conv, C3 and SPP. We choose YOLOv5s as the detection model. Firstly, the original 416 × 416 × 3 image is input into Focus model and segmented into four slices with the size of 208 × 208 × 3 per slice by using a slicing operation. Secondly, the four slices are connected to a feature map of size 208 × 208 × 12 using the concat operation, and finally form a feature map of 208 × 208 × 32 (Fig. 3) through the convolution operation of 32 convolution cores (Fig. 3). C3 model contains three convolutions and is used to extract the deep features of the image. The SPP is used to concatenate feature maps from different kernel sizes (e.g., 13×13, 9×9, 5×5) together as an output, effectively increasing the receptive field of the backbone network and separating significant context features (He et al., 2015).

(a) Cow dataset samples

(b) Beef dataset samples

Fig. 1. Examples of cattle dataset samples.

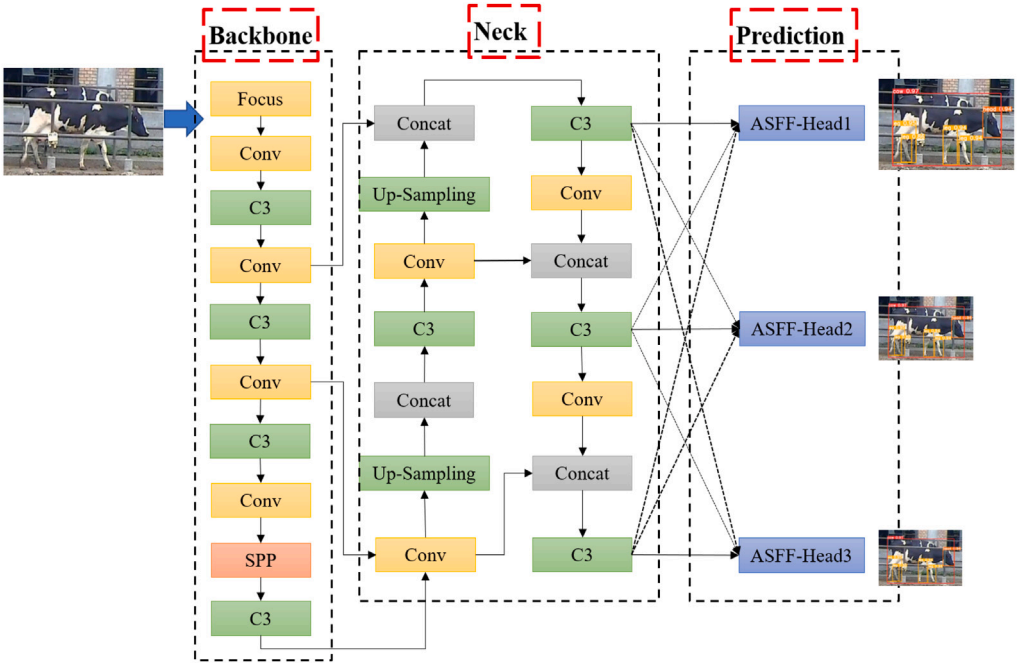


Fig. 2. The overall structure of YOLOv5-ASFF network.

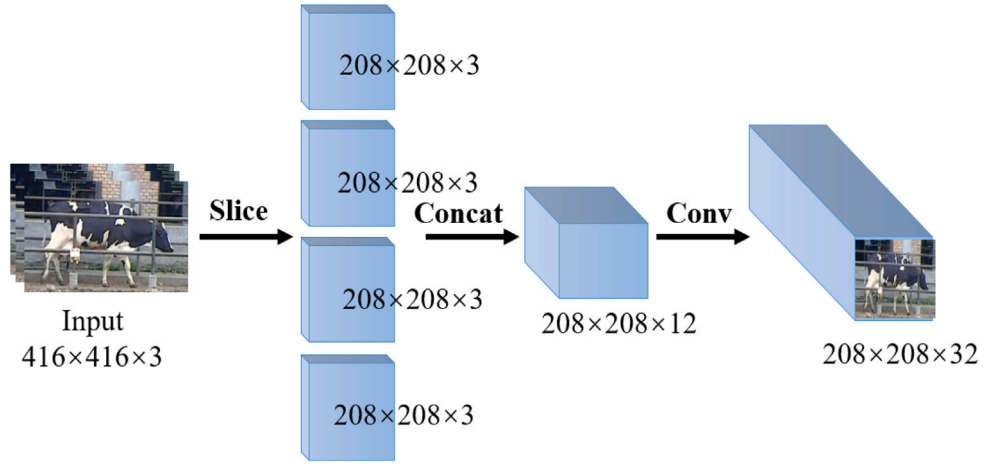


Fig. 3. Structure of Focus module.

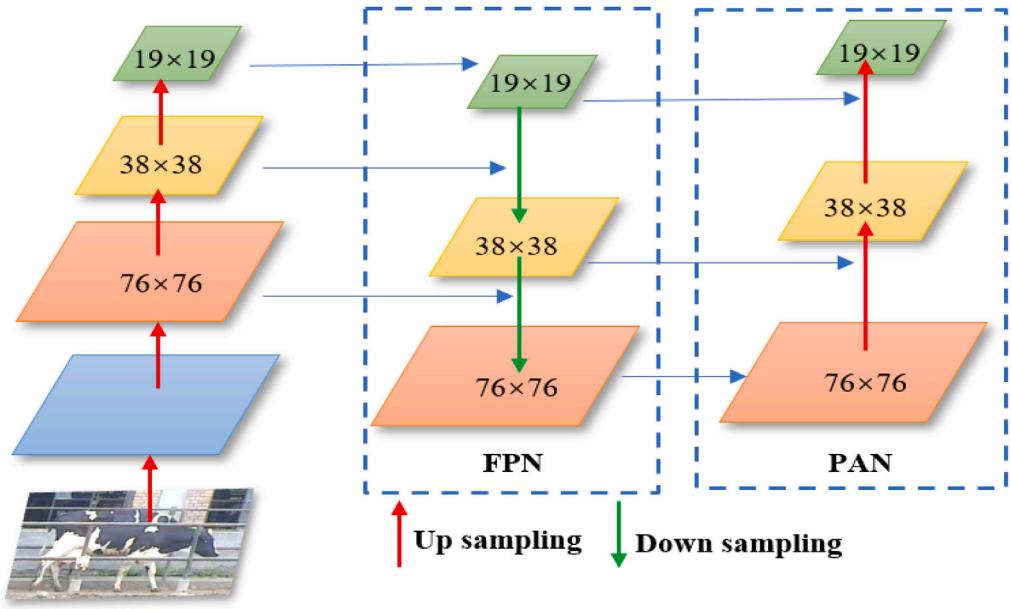


Fig. 4. FPN-PAN structure.

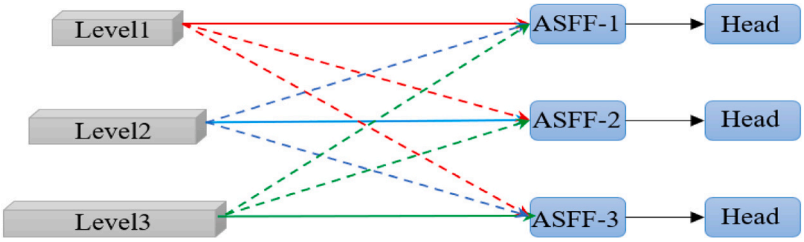


Fig. 5. The structure of ASFF.

2.2.2. Neck of YOLOv5-ASFF model

The Neck of YOLOv5-ASFF model is used to achieve the aggregation of semantic features at three scales to identify cattle objects of different sizes. As shown in Fig. 4, the feature maps of three scales extracted by Backbone are sent to the FPN-PAN structure of the Neck for feature aggregation, which is beneficial to improve the accuracy of cattle target detection. The FPN-PAN structure consists of a bottom-up path and a top-down path, an effective multi-scale feature fusion method, which is mainly used to generate feature pyramids, enhance the model's detection of objects of different scales, and realize the recognition of the same object of different sizes and scales (Yao et al., 2021). In addition, C3 modules are also added at this stage to enhance the feature fusion capability.

2.2.3. ASFF-head of YOLOv5-ASFF model

The background of farm images is complex and constantly changing, which introduces considerable background noise. Therefore, the cattle detection network requires a powerful feature-fusion module. However, the FPN-PAN structure has the problem of insufficient fusion of multi-scale features, so we introduce ASFF module into the YOLOv5s head to autonomously learn the spatial weight of each scale to achieve full fusion of image multi-scale features. The ASFF module can filter features at three levels after neck and combine the helpful information (Feng and Yi, 2022; Liu et al., 2019). The ASFF structure is shown in Fig. 5.

Three feature maps level1, level2 and level3 are obtained through the neck of Yolov5-ASFF. As shown in Fig. 5, take ASFF-1 as an example. The fused ASFF-1 output is the result of multiplying and adding the semantic features of level1, level2, and level3 with weights α, β and γ from different layers (Zhang et al., 2021). As Eq. (1) shows:

$$ASFF-1 = \alpha^1 \cdot x^{1\to1} + \beta^1 \cdot x^{2\to1} + \gamma^1 \cdot x^{3\to1} \tag{1}$$

where, x^1, x^2 and x^3 represent the characteristic diagrams from Level1, Level2 and Level3 respectively. $x^{2\to1}$ means that the size of the feature map of Level2 is adjusted to the size of Level1. Similarly, $x^{3\to1}$ is the same meaning. In the process of feature fusion, the adjusted features are multiplied by their corresponding weight parameters α^1, β^1 and γ^1 respectively, and then the results are added to obtain new fusion features, which are the output feature map of the ASFF-1 network module.

Let $x_{ij}^{n\to l}$ represents the feature vector at the position (i, j) on the feature maps from level N to level L. For feature fusion of level L, the process is shown in Eq. (2):

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1\to l} + \beta_{ij}^l \cdot x_{ij}^{2\to l} + \gamma_{ij}^l \cdot x_{ij}^{3\to l} \tag{2}$$

where, y_{ij}^l represents the vector at position (i, j) of the output feature maps y^l among channels. α_{ij}^l, β_{ij}^l and γ_{ij}^l represent the spatial importance weights of the feature map at three different levels to the level L, which are adaptively learned by the network. Note that α_{ij}^1, β_{ij}^1 and γ_{ij}^1 are scalar variables which are shared across all channels. $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$

and α_{ij}^{l}, β_{ij}^{l}, $\gamma_{ij}^{l} \in [0, 1]$, and define:

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}} + e^{\lambda_{\beta_{ij}}^{l}} + e^{\lambda_{\gamma_{ij}}^{l}}} \tag{3}$$

where α_{ij}^{l}, β_{ij}^{l} and γ_{ij}^{l} are defined by $\lambda_{\alpha_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$ and $\lambda_{\gamma_{ij}}^{l}$ softmax function as the control parameters respectively. We use (1 × 1) convolutional layer to calculate the weights of scalar maps, which can be learned by standard backpropagation, the features of each level are adaptively aggregated at each scale, and the output is used to detect the transport path that follows YOLOv5. After that, the fused feature map is input into head for classification and identification of cattle individual, head, and legs.

2.2.4. Loss function

The loss function based on YOLOv5-ASFF model is composed of objectness loss (l_{obj}), classification loss (l_{cls}) and location loss function (l_{CIoU}), which was used in the cattle detection model based on YOLOv5-ASFF. The loss function equations are as follows (Wang et al., 2021a):

$$Loss = l_{obj} + l_{cls} + l_{CIoU} \tag{4}$$

$$\begin{aligned} l_{obj} = &- \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{obj} [\hat{C}_i log(C_i) + (1 - \hat{C}_i) log(1 - C_i)] - \\ &\lambda_{noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{noobj} [\hat{C}_i log(C_i) + (1 - \hat{C}_i) log(1 - C_i)] \end{aligned} \tag{5}$$

$$\begin{aligned} l_{cls} = &- \sum_{i=0}^{S^2} I_{i,j}^{obj} \sum_{j=0} [\hat{P}_i(c) log(P_i(c)) \\ &+ (1 - \hat{P}_i(c)) log(1 - P_i(c))] \end{aligned} \tag{6}$$

$$l_{CIoU} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{i,j}^{obj} [1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v] \tag{7}$$

$$v = \frac{4}{\pi^2} (arctan \frac{w^{gt}}{h^{gt}} - arctan \frac{w}{h})^2 \tag{8}$$

$$\alpha = \frac{v}{(1 - IoU) - v} \tag{9}$$

where, S is the number of grids, b is the number of prior boxes in each grid. λ_{noobj} represents weight. $I_{i,j}^{obj}$ and $I_{i,j}^{noobj}$ are used to determine whether the jth priori box of the ith grid contains the object. If yes, $I_{i,j}^{obj}$ is 1 and $I_{i,j}^{noobj}$ is 0. Otherwise $I_{i,j}^{obj}$ is 0 and $I_{i,j}^{noobj}$ is 1, only the confidence loss of the box is calculated. $\hat{C}_i$ indicates whether the anchor frame is responsible for the prediction of the whole network. If yes, it is 1, otherwise it is 0. C_i means the confidence levels of prediction and tagging box. $P_i(c)$ is the probability that the current prior box has the target region. c is the diagonal distance of the minimum closure region containing the prediction box and the actual box. IoU means the ratio of the intersection and union of the prediction bounding box and the actual bounding box. $\rho(\cdot)$ is the Euclidean distance. b, w and h are the center coordinates, width and height of the prediction box respectively. b^{gt}, w^{gt} and h^{gt} are the center coordinates, width and height of the actual box respectively. α is weight factor, v is similarity ratio of length to width.

2.3. Network training parameters

In this work, all experiments were conducted on a computer equipped with a GeForce GTX 1080 Ti GPU, I9-7920X CPU@2.9 GHz. In order to compare these methods fairly, all the network's input sizes were set to 416 × 416 × 3, the training epoch was set to 1000, batch size was 16, and the learning rate was 0.0013. In addition, all the network's initial weights were random. The other parameters of each network were their default settings.

To verify the effectiveness of the proposed approach, the proposed YOLOv5-ASFF was compared with other state-of-the-art methods—Faster R-CNN (Ren et al., 2015),Cascade R-CNN (Cai and Vasconcelos, 2018), SSD (Liu et al., 2016), YOLOv3 (Redmon and Farhadi, 2018), YOLOv5s (Liu et al., 2021b).

2.4. Performance evaluation

In our work, precision, recall, mean average precision (mAP) and Frames Per Second (FPS) were used to evaluate the performance of YOLOv5-ASFF. And the IOU (Intersection Over Union) of the predicted frame and the actual marked frame is used to judge whether the target has been successfully predicted. When the IOU ≥ 0.5, the target can be successfully predicted; When the IOU < 0.5, the target prediction is wrong. TP, FP and FN are the numbers of true positive samples, false positive samples and false negative samples, respectively. The relevant evaluation indicators as follows:

(1) Precision. Precision indicates the proportion of the number of correct predictions in the identified images in the prediction results.

$$Precision = \frac{tp}{tp + fp} \times 100\% \tag{10}$$

(2) Recall. Recall represents the proportion of the number of correct predictions in all samples of the test set.

$$Recall = \frac{tp}{tp + fn} \times 100\% \tag{11}$$

(3) F1. The F1-score is the reconciled mean of precision and recall.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \times 100\% \tag{12}$$

(4) mAP. The mAP is the average value of the AP (Average Precision), when the values were higher, the detection effect of the algorithm was better.

$$AP = \int_0^1 P(r) dr \tag{13}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP(i) \tag{14}$$

where n denotes the number of classification types (n is 4).

3. Results and discussion

Datasets in this study included cow datasets and beef datasets in different scenes (e.g. day, night, single, multiple) which were used to test the applicability and effectiveness of YOLOv5-ASFF. In this paper, Faster R-CNN, Cascade R-CNN, SSD, YOLOv3 and YOLOv5s were selected for comparative analysis.

3.1. Results of detection models

The detection results of the key parts of cattle with different models are shown in Table 1. From Table 1, we can see that the precision, recall, F_1 and mAP@0.5 of the YOLOv5-ASFF was up to 96.2%, 92%, 94.1% and 94.7%, which was higher than that of YOLOv5s (95.8%, 91.4%, 93.5% and 94%), followed by Faster R-CNN (87.2%, 94.5%, 90.7% and 90.2%) and Cascade R-CNN (88.7%, 94.8%, 91.6% and 90.3%). And the results of SSD (58.1%, 92.6%, 71.4% and 85.4%) and YOLOv3 (61.3%, 92.3%, 73.7% and 85.7%) were worse. The results show that the overall performance of the proposed YOLOv5-ASFF is the best. In addition, it can be seen from Fig. 1 that the scene of the sample dataset is complex, and the scene contains many small target cattle, which may lead to the reason for the low precision of SSD and YOLOv3.

In addition, the detection speed of our proposed YOLOv5-ASFF is 41 FPS, which is lower than that of SSD (50 FPS) and YOLOv3 (60 FPS),

Table 1
Performance comparison of different algorithms (%).

Methods	Precision (%)	Recall (%)	F_i (%)	mAP@0.5 (%)	FPS(Frame/s)
Fass R-CNN	87.2	94.5	90.7	90.2	16
Cascade R-CNN	88.7	94.8	91.6	90.3	13
SSD	58.1	92.6	71.4	85.4	50
YOLOv3	61.3	92.3	73.7	85.7	60
YOLOv5s	95.8	91.4	93.5	94	58
YOLOv5-ASFF	96.2	92	94.1	94.7	41

Fig. 6. Cattle detection results of YOLOv5-ASFF.

but the precision rate is much higher than that of SSD and YOLOv3. And 41 FPS meets the real-time requirement (>20 FPS) of animal detection in precision agriculture. Overall, the proposed YOLOv5-ASFF enhances the animal by learning fused spatial weights for each scale feature map, which provides a favorable solution for the remote animal detection in smart livestock farming.

To verify the generalization performance of the proposed YOLOv5-ASFF, cattle (cow and beef) images containing multiple targets from different time periods were selected for testing. Fig. 6 shows the detection results of our proposed approach. It can be seen that the YOLOv5-ASFF model can detect cow, beef, head and legs well with different lighting, occlusion (foreign object occlusion, occlusion between animals), multiple targets and small targets. All this demonstrate the robustness of the proposed YOLOv5-ASFF approach.

3.2. Test results of key cattle body parts

To further analysis the detection performance of different body parts, Table 2 lists the precision rates of different methods for body parts detection. It can be seen from Table 2 that all the precision rates of cows are higher than that of beef, because the distance between the cow and the camera in the cow dataset is basically unchanged, while the distance between the beef and the camera in the beef dataset is inconsistent. That is, the inconsistent depth of field leads to the existence of large and small objects in the scene, so the detection precision of beef is low. In addition, it also shows that SSD and YOLOv3 have poor application effect in the detection of both large and small objects in the same scene.

In addition, the detection results of Table 2 show that the precision of the YOLOv5-ASFF for cow (98.5%), beef (93.3%), head (96.8%) and leg (96.2%) is higher than that of YOLOv5s (98.2% for cow, 92.8% for beef, 96.7% for head, and 95.5% for legs). Fig. 7 shows detection results using YOLOv5s and YOLOv5-ASFF in cattle dataset. It can be seen that due to the influence of occlusion or depth of field, YOLOv5s will falsely detect or miss detection during detection. However, YOLOV5-ASFF has improved detection performance in the above cases.

In order to further compare YOLOv5s and YOLOv5-ASFF, the PR curve for each parts of the cattle was obtained (Fig. 8). As can be seen from Fig. 8, for the cattle body classification (cow, beef, head, legs), the detection accuracy of YOLOv5-ASFF algorithm is higher than that of YOLOv5s in all cattle parts. It can be seen that the introduction of ASFF in YOLOv5 can indeed improve the overall performance.

3.3. Performance analysis of detection performance

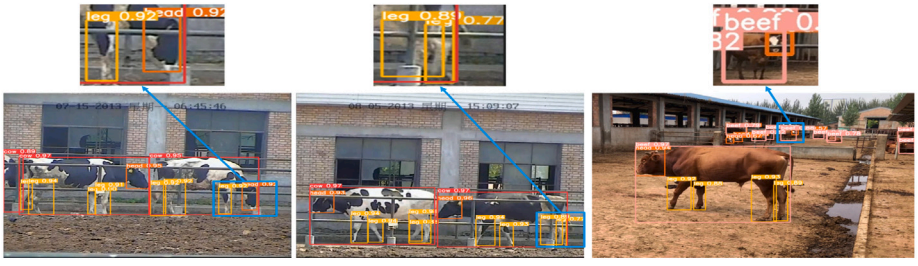
From the results, we can see the YOLOv5-ASFF approach outperforms Faster R-CNN, Cascade R-CNN, SSD, YOLOv3 and YOLOv5s in the detection of cattle dataset. The precision rates of YOLOv5-ASFF and YOLOv5s are 96.2% and 95.8%, which has the best detection effect. But it can be seen from Table 2 that YOLOv5-ASFF is better than YOLOv5s in the detection of cattle body parts, especially the legs (96.2% for YOLOv5-ASFF, 95.5% for YOLOv5-ASFF).

In addition, due to the differences in the R-CNN, SSD and YOLOv3 network structures, the detection results are quite different. R-CNN uses the Selective Search method to pre-extract a series of candidate regions that are more likely to be objects, and then uses CNN to extract features on these candidate regions for judgment. Although the detection results are good, the efficiency is slow (Table 1). The SSD model is very sensitive to the size of the bounding box. That is to say, SSD is more sensitive to small object targets and performs poorly in detecting small object targets (Liu et al., 2016), so it does not perform well on this dataset. YOLO v3 combines some of the advantages of R-CNN and SSD,

Table 2
Comparison of different body part detection performance.

Methods	Cow precision (%)	Beef precision (%)	Heads precision (%)	Legs precision (%)
Fass R-CNN	92.9	78.1	90.8	86.8
Cascade R-CNN	96.3	79.8	90.5	88.1
SSD	84.1	42.6	49.5	56
YOLOv3	79.4	46.3	52.5	66.8
YOLOv5s	98.2	92.8	96.7	95.5
YOLOv5-ASFF	98.5	93.3	96.8	96.2

(a) Examples of YOLOv5 detection results



(b) Examples of YOLOv5-ASFF detection results

Fig. 7. Examples of cattle detection results using YOLOv5 and YOLOv5-ASFF.

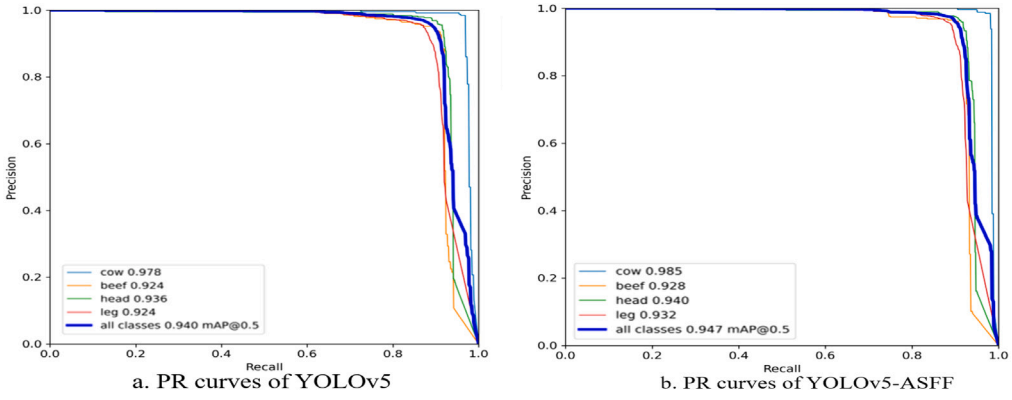


Fig. 8. PR curves of YOLOv5s and YOLOv5-ASFF.

which is faster than R-CNN and more accurate than SSD for detecting small targets (Tables 1 and 2). However, the overall detection effect on this dataset is still not good. With the YOLO family of methods, YOLOv5 has been greatly improved in object detection, and has solved the above problems well (Zhu et al., 2021; Zhao et al., 2021).

In summary, YOLOv5-ASFF has achieved good results in large and small targets detection, and can be applied to animal detection in different scenes.

3.4. Data analysis of cattle samples

The cattle data samples contain cow data and beef data. The selected scenes include different lighting, occlusion (foreign object occlusion, occlusion between animals), different depths of field, multiple targets, etc. Which are used to construct a sample dataset in a real breeding environment (Fig. 1). In this dataset, YOLOv5s and YOLOv5-ASFF have the best detection results. However, due to the existence of occlusion and small target area in sample data, some information is

lost, and thus the feature learning is insufficient, so YOLOv5 has false detection or missed detection (Fig. 7a). In addition, for small targets, after multi-layer convolution, there is basically not much information left. ASFF model is to adaptively learn the spatial weights of the fusion of feature maps at various scales, which can learn the target features more effectively. Therefore, combining ASFF and YOLOv5 can improve the overall recognition accuracy, especially in small target areas (Fig. 7b).

3.5. Model application analysis

This study realized the detection of individual and key body parts (head and legs) of cow and beef in different scenes, which showed the robustness and applicability of this method. YOLOv5-ASFF is small in size and can be transplanted to the mobile platform, which can be used as portable detection, so it is more flexible in application and greatly facilitates the monitoring of farm information. The detection performance of the proposed method at night still needs further research. The contour of animals at night is not obvious, and a large number of body surface information is missing, which reduces the detection accuracy. However, the nocturnal animal information can be highlighted through image enhancement, such as GAN network enhancement (Lu et al., 2022). On the basis of the detected key body parts, each part can be further tracked, and the motion information of each part can be correlated to more accurately identify animal behavior.

4. Conclusion

In order to realize the detection of cattle in different feeding scenes, a cattle body parts detection method based on the YOLOv5-ASFF model was proposed in this study. The proposed approach integrating adaptively spatial feature fusion (ASFF) into YOLOv5 fully extracts and learns the target features, and achieves good detection performance in both cow and beef datasets, which provides a new method for longterm and real-time animal detection in smart livestock farming. The proposed YOLOv5-ASFF based approach achieved a precision of 96.2%, a recall of 92%, an F1 score of 94.1%, and an mAP@0.5 of 94.7%, which outperformed Faster R-CNN, Cascade R-CNN, SSD, YOLOv3 and YOLOv5s. In addition, YOLOv5-ASFF achieves 98.5%, 93.3%, 96.8% and 96.2% detection precisions for cow, beef, head and legs, respectively, which are all higher than those of the comparison methods. Experimental results demonstrated that the proposed YOLOv5-ASFF could improve the performance of cattle detection in complex environments. And the detection of head and legs can further obtain local information of cattle, which is conducive to accurate monitoring. However, the detection effect is still not ideal for dense, severely occluded, night images and other data sets that cause a lot of information loss. In future work, animal information will be improved by collecting images from multiple perspectives, and feature information will be highlighted by image enhancement methods. In addition, the YOLOv5-ASFF method can be further embedded into the robotic platform to achieve automatic and efficient cattle or key body parts detection. Then the movement information of the head and legs can be further obtained, which is conducive to the analysis of animal behavior.

CRediT authorship contribution statement

Yongliang Qiao: Came up the research thoughts, Figured out the research methodology, Developed the algorithm for data analysis, Analyzed the data, Verified the results, Wrote the manuscript. **Yangyang Guo:** Came up the research thoughts, Figured out the research methodology, Developed the algorithm for data analysis, Analyzed the data, Verified the results, Wrote the manuscript. **Dongjian He:** Came up the research thoughts, Figured out the research methodology, Wrote the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

The authors particular thanks to other members of the team for their involvement and efforts in the whole experiment organization and information collection. The authors also acknowledge the support by the project: National Natural Science Foundation of China (grant number 61473235) and the National Key Technology R&D Program of China (grant number 2017YFD0701603).

References

Beggs, D., Jongman, E., Hemsworth, P., Fisher, A., 2019. Lame cows on Australian dairy farms: A comparison of farmer-identified lameness and formal lameness scoring, and the position of lame cows within the milking order. J. Dairy Sci. 102 (2), 1522–1529.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Feng, J., Yi, C., 2022. Lightweight detection network for arbitrary-oriented vehicles in UAV imagery via global attentive relation and multi-path fusion. Drones 6 (5), 108.

Fournel, S., Rousseau, A.N., Laberge, B., 2017. Rethinking environment control strategy of confined animal housing systems through precision livestock farming. Biosyst. Eng. 155, 96–123.

Gertz, M., Große-Butenuth, K., Junge, W., Maassen-Francke, B., Renner, C., Sparenberg, H., Krieter, J., 2020. Using the XGBoost algorithm to classify neck and leg activity sensor data using on-farm health recordings for locomotor-associated diseases. Comput. Electron. Agric. 173, 105404.

Guo, Y., Qiao, Y., Sukkarieh, S., Chai, L., He, D., 2021. Bigru-attention based cow behavior classification using video data for precision livestock farming. Trans. ASABE 64 (6), 1823–1833.

Guo, Y., Zhang, Z., He, D., Niu, J., Tan, Y., 2019. Detection of cow mounting behavior using region geometry and optical flow characteristics. Comput. Electron. Agric. 163, 104828.

He, D., Liu, D., Zhao, K., 2016. Review of perceiving animal information and behavior in precision livestock farming. Trans. Chin. Soc. Agric. Mach. 47 (5), 231–244.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37 (9), 1904–1916.

Jiang, B., Wu, Q., Yin, X., Wu, D., Song, H., He, D., 2019. FLYOLOv3 deep learning for key parts of dairy cow body detection. Comput. Electron. Agric. 166, 104982.

Li, X., Cai, C., Zhang, R., Ju, L., He, J., 2019. Deep cascaded convolutional models for cattle pose estimation. Comput. Electron. Agric. 164, 104885.

Lim, J., Pyo, S., Kim, N., Lee, J., Lee, J., 2020. Obstacle magnification for 2-D collision and occlusion avoidance of autonomous multirotor aerial vehicles. IEEE/ASME Trans. Mechatronics 25 (5), 2428–2436.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2117–2125.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C., 2016. Ssd: Single shot multibox detector. In: European Conference on Computer Vision. Springer, pp. 21–37.

Liu, S., Cheng, J., Liang, L., Bai, H., Dang, W., 2021a. Light-weight semantic segmentation network for UAV remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 8287–8296.

Liu, S., Huang, D., Wang, Y., 2019. Learning spatial fusion for single-shot object detection. arXiv preprint arXiv:1911.09516.

Liu, S., Qi, L., Qin, H., Shi, J., Jia, J., 2018. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8759–8768.

Liu, K., Tang, H., He, S., Yu, Q., Xiong, Y., Wang, N., 2021b. Performance validation of YOLO variants for object detection. In: Proceedings of the 2021 International Conference on Bioinformatics and Intelligent Computing. pp. 239–243.

Lu, Y., Chen, D., Olaniyi, E., Huang, Y., 2022. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. Comput. Electron. Agric. 200, 107208.

Mishra, S., Syed, D.F., Ploughe, M., Zhang, W., 2021. Autonomous vision-guided object collection from water surfaces with a customized multirotor. IEEE/ASME Trans. Mechatronics 26 (4), 1914–1922.

Patil, R.K., Patil, S.S., 2020. Cognitive intelligence of Internet of Things in smart agriculture applications. In: 2020 IEEE Pune Section International Conference. PuneCon, IEEE, pp. 129–132.

Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Yoshioka, H., Itoyama, E., et al., 2019. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. Comput. Electron. Agric. 157, 247–253.

Qiao, Y., Cappelle, C., Ruichek, Y., Yang, T., 2019a. ConvNet and LSH-based visual localization using localized sequence matching. Sensors 19 (11), 2439.

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.

Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., Sukkarieh, S., 2021. Intelligent perception for cattle monitoring: A review for cattle identification, body condition score evaluation, and weight estimation. Comput. Electron. Agric. 185, 106143.

Qiao, Y., Truman, M., Sukkarieh, S., 2019b. Cattle segmentation and contour extraction based on mask R-CNN for precision livestock farming. Comput. Electron. Agric. 165, 104958.

Redmon, J., Farhadi, A., 2018. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

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.

Riekert, M., Klein, A., Adrion, F., Hoffmann, C., Gallmann, E., 2020. Automatically detecting pig position and posture by 2D camera imaging and deep learning. Comput. Electron. Agric. 174, 105391.

Shafiee, M.J., Chywl, B., Li, F., Wong, A., 2017. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. arXiv preprint arXiv:1709.05943.

Shao, W., Kawakami, R., Yoshihashi, R., You, S., Kawase, H., Naemura, T., 2020. Cattle detection and counting in UAV images based on convolutional neural networks. Int. J. Remote Sens. 41 (1), 31–52.

Sun, G., Shi, C., Liu, J., Ma, P., Ma, J., 2021. Behavior recognition and maternal ability evaluation for Sows based on triaxial acceleration and video sensors. IEEE Access 9, 65346–65360.

Wang, D., He, D., 2021. Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. Biosyst. Eng. 210, 271–281.

Wang, C., Luo, Q., Chen, X., Yi, B., Wang, H., 2021a. Citrus recognition based on YOLOv4 neural network. 1820, (1), IOP Publishing, 012163,

Wang, K., Wu, P., Cui, H., Xuan, C., Su, H., 2021b. Identification and classification for sheep foraging behavior based on acoustic signal and deep learning. Comput. Electron. Agric. 187, 106275.

Yan, B., Fan, P., Lei, X., Liu, Z., Yang, F., 2021. A real-time apple targets detection method for picking robot based on improved YOLOv5. Remote Sens. 13 (9), 1619.

Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X., 2021. A real-time detection algorithm for Kiwifruit defects based on YOLOv5. Electronics 10 (14), 1711.

Zhang, Z., Kayacan, E., Thompson, B., Chowdhary, G., 2020. High precision control and deep learning-based corn stand counting algorithms for agricultural robot. Auton. Robots 44 (7), 1289–1302.

Zhang, R., Shi, Y., Yu, X., 2021. Pavement crack detection based on deep learning. In: 2021 33rd Chinese Control and Decision Conference. CCDC, IEEE, pp. 7367–7372.

Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., Zhu, Y., Cao, W., 2021. A wheat spike detection method in UAV images based on improved YOLOv5. Remote Sens. 13 (16), 3095.

Zhu, L., Geng, X., Li, Z., Liu, C., 2021. Improving yolov5 with attention mechanism for detecting boulders from planetary images. Remote Sens. 13 (18), 3776.