Article

A Study on Multi-Scale Behavior Recognition of Dairy Cows in Complex Background Based on Improved YOLOv5

Zheying Zong [1,2,3,4], Zeyu Ban [1] , Chunguang Wang [1], Shuai Wang [1], Wenbo Yuan [1], Chunhui Zhang [1,*], Lide Su [1] and Ze Yuan [1]

[1] College of Electromechanical Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China; zongzheying@imau.edu.cn (Z.Z.); 2022202050027@emails.imau.edu.cn (Z.B.); jdwcg@imau.edu.cn (C.W.); wangshuai@imau.edu.cn (S.W.); ywb@emails.imau.edu.cn (W.Y.); sld@imau.edu.cn (L.S.); nmgnydxyz@emails.imau.edu.cn (Z.Y.)
[2] The Innovation Team of Higher Education Institutions in Inner Mongolia Autonomous Region, Hohhot 010018, China
[3] Inner Mongolia Engineering Research Center of Intelligent Equipment for the Entire Process of Forage and Feed Production, Hohhot 010018, China
[4] Full Mechanization Research Base of Dairy Farming Engineering and Equipment, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Hohhot 010018, China
* Correspondence: zhangchunhui@imau.edu.cn

Abstract: The daily behaviors of dairy cows, including standing, drinking, eating, and lying down, are closely associated with their physical health. Efficient and accurate recognition of dairy cow behaviors is crucial for timely monitoring of their health status and enhancing the economic efficiency of farms. To address the challenges posed by complex scenarios and significant variations in target scales in dairy cow behavior recognition within group farming environments, this study proposes an enhanced recognition method based on YOLOv5. Four Shuffle Attention (SA) modules are integrated into the upsampling and downsampling processes of the YOLOv5 model's neck network to enhance deep feature extraction of small-scale cow targets and focus on feature information, while maintaining network complexity and real-time performance. The C3 module of the model was enhanced by incorporating Deformable convolution (DCNv3), which improves the accuracy of cow behavior characteristic identification. Finally, the original detection head was replaced with a Dynamic Detection Head (DyHead) to improve the efficiency and accuracy of cow behavior detection across different scales in complex environments. An experimental dataset comprising complex backgrounds, multiple behavior categories, and multi-scale targets was constructed for comprehensive validation. The experimental results demonstrate that the improved YOLOv5 model achieved a mean Average Precision (mAP) of 97.7%, representing a 3.7% improvement over the original YOLOv5 model. Moreover, it outperformed comparison models, including YOLOv4, YOLOv3, and Faster R-CNN, in complex background scenarios, multi-scale behavior detection, and behavior type discrimination. Ablation experiments further validate the effectiveness of the SA, DCNv3, and DyHead modules. The research findings offer a valuable reference for real-time monitoring of cow behavior in complex environments throughout the day.

Keywords: dairy cow; behavior recognition; improved YOLOv5; Shuffle Attention; deformable convolution; Dynamic Detection Head

1. Introduction

The behavior of dairy cows, influenced by internal physiological changes or external environmental stimuli, serves as a direct or indirect indicator of their health and physiolog-

ical condition. It plays a critical role in monitoring diseases and detecting abnormalities in dairy cows [1–3]. The daily behaviors of dairy cows encompass walking, lying down, drinking, eating, and others. For instance, lying down is one of the most essential behaviors in the daily activities of dairy cows. Typically, dairy cows require 10 to 14 h of lying down per day. This behavior is not only vital for their physical health but also directly correlates with milk production. Studies indicate that each additional hour of lying down time can increase milk production by approximately 1.7 kg. A reduction in lying time may result from factors such as uncomfortable bedding or estrus [4,5].

In modern dairy farming, traditional manual observation methods are inadequate for refined behavior monitoring due to their inefficiency and lack of timeliness. Accurate understanding of the unique living habits and behavioral characteristics of cows is crucial for the prevention, diagnosis, and treatment of diseases [6–8]. Therefore, automation, intelligence, and precision have become inevitable trends in large-scale, intensive dairy farming to meet its developmental needs [9–11]. With the rapid advancement of computer vision technology, numerous scholars have initiated research on cow behavior monitoring methods using video image analysis. Early methods for monitoring cow behavior depended on image or video feature extraction combined with traditional classifiers, which significantly contributed to initial research on cow behavior analysis [12,13]. Related studies have achieved high recognition rates for single-behavior tasks through specific feature extraction algorithms, including background subtraction and optical flow analysis [14,15]. He et al. [16] achieved high-precision detection of behaviors such as lying and standing in calves using a maximum connected region target cyclic search algorithm, achieving a 100% recognition rate for certain behaviors. Additionally, Song et al. [17] employed optical flow algorithms to observe the cow's abdominal movement during breathing, achieving accurate detection with an average accuracy of 98.58%. However, these methods heavily rely on artificial feature design for specific behaviors, are sensitive to background noise in complex scenarios, and exhibit limited generalization capabilities. Their performance is limited in multi-behavior detection tasks, failing to meet the complex requirements of practical livestock farm applications.

With the rapid advancement of deep learning technology, devices such as cameras can now capture RGB images to simulate biological vision, enabling target detection and animal behavior recognition [18–21]. Deep learning technology employs convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) to automatically learn key features from extensive data, thereby minimizing reliance on artificial feature design [22–25]. For single-behavior identification, Zhang et al. [26] utilized the YOLOv3 model to detect beef cattle targets, determine their positions, and identify feeding behavior through a convolutional neural network, enabling the recognition of feeding behavior in multiple targets. Liu et al. [27] enhanced the structure and parameter design of the convolutional neural network to achieve efficient detection of estrus climbing behavior in cows within a single background, achieving an accuracy of 98.25%. Wang et al. [28] optimized the YOLOv3 network, achieving a 99% recognition rate for single climbing behavior at multiple scales but only 86.27% for long-distance and small-scale climbing behavior. For multi-behavior target detection, Yin et al. [29] employed a bidirectional feature pyramid network and a bidirectional long short-term memory method to distinguish between natural behaviors such as lying, standing, and walking, achieving an accuracy of 97.87%. Ma et al. [30] extended the RexNet 3D network to a three-dimensional algorithm, utilizing the SoftMax function to classify lying, standing, and walking behaviors in complex backgrounds, with experimental results demonstrating an accuracy of 95%. Video-based cow behavior recognition methods exhibit high accuracy, but the datasets used often feature simple environments with single targets, resulting in poor multi-scale behavior

recognition. These limitations hinder their ability to meet the practical requirements of real-time monitoring in complex environments. Therefore, further optimization and improvement of the algorithm are necessary for practical applications.

This study aims to enhance the accuracy of cow behavior recognition and detection in complex environments, mitigate the impact of multi-scale targets on recognition results, and improve the model's performance in identifying multiple behaviors. This study introduces SA [31], which integrates channel attention and spatial attention mechanisms to enhance the model's focus on behavior regions, effectively suppress complex background interference, and strengthen deep feature extraction and attention to small-scale targets. The C3 module is enhanced by incorporating DCNv3 [32], enabling the model to adapt to variations in target scale and shape complexity and improving the accuracy of cow behavior characteristic identification. Finally, to improve detection head performance, the original detection head was replaced with DyHead [33], enabling efficient and accurate detection of cow behavior across different scales in complex environments. The YOLOv5 model was optimized to enhance its multi-scale target detection capability in complex environments, aiming to improve cow behavior recognition.

2. Materials and Methods

2.1. Data Source

This experiment focused on Chinese Holstein cows and was conducted at a dairy farm located in Shangshilipo New Village, Tumotezuoqi, Hohhot City, Inner Mongolia Autonomous Region (111°33′ E, 40°74′ N, 1058 m above sea level). A plan view of the dairy cow test site is provided in Figure 1.
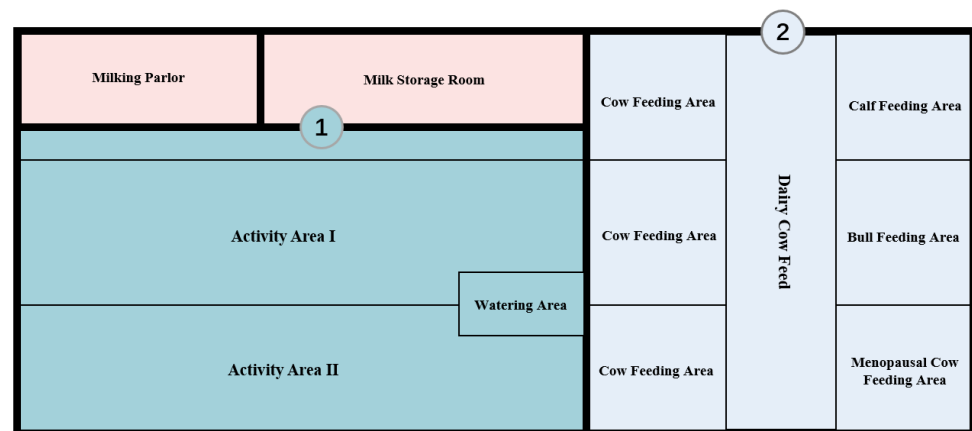
Figure 1. Plan of the dairy cow test site. Note: Camera 1 (Dahua P40A20-WT-1) captures the outdoor activity areas I and II of the cows; camera 2 (Xiaomi Smart Camera 3 Pan and Tilt Version) captures the indoor feeding area of the cows.

Two network cameras, the Dahua P40A20-WT-1 (Da Hua Technology, Hangzhou, China) and the Xiaomi Smart Camera 3 Pan and Tilt Edition (Xiaomi Technology, Shanghai, China), were employed to film the cow activity and feeding areas, capturing behaviors such as eating, drinking, standing, and lying down. The cow feeding area covers approximately 200 m^2. The second camera was mounted on the indoor wall, 3.5 m above the ground, at the center of the area, capturing the cow feeding area from a bird's-eye view to record indoor feeding behavior. A schematic diagram of the camera installation in the indoor feeding area is presented in Figure 2a. The cow activity area spans approximately 600 m^2. The camera was installed on the outer wall of the milking room, 2.8 m above the ground, at the center of the field, capturing the cow activity area from a bird's-eye view to record outdoor

standing, drinking, and lying behaviors. A schematic diagram of the camera installation in the activity area is illustrated in Figure 2b.

Figure 2. Dairy cow test site. (**a**) Diagram of camera in cow feeding area; (**b**) diagram of cow activity area camera.

2.2. Datasets and Models

2.2.1. Dataset Construction

To evaluate the recognition performance of the improved YOLOv5 network, real surveillance videos from actual housing environments were utilized as test data. To assess the model's recognition capability across different scenarios, five video clips from diverse environments were selected for testing, labeled as 01 to 05. Based on manual observation, videos with a high density of cows and significant occlusion due to overlapping postures were classified as dense, while those with fewer cows and minimal occlusion were classified as sparse. Details are presented in Table 1. Notably, the number of cows in the activity area is generally higher during the morning (8:00–12:00) and afternoon (12:00–17:00) periods compared to the sunrise (5:00–7:00) and evening (18:00–21:00) periods, with a higher probability of occlusion among cows. From the perspective of cow activity, mornings and afternoons are periods of increased activity and higher likelihood of gathering. Cow behavior during these periods is highly diverse. Data collection during these periods captured cows in various states, including standing, eating with heads down, drinking, and other postures and movements.

Table 1. Dairy cow target recognition dataset.

Sequence	Weather	Period	Sparse	Dense	Interference Factors
01	cloudy	sunrise	√	-	light: weak shading: slight
02	sunny	morning	-	√	light: normal shading: moderate
03	sunny	afternoon	-	√	lighting: strong shade: heavy
04	cloudy	evening	-	√	light: weak shading: moderate
05	cloudy	night	√	-	lighting: dark shading: slight

This study involved 38 Holstein cows as research subjects. Video footage of cow behaviors was collected from April to December 2023. A total of 1044 videos of cow behaviors were selected, comprising 1008 short videos of approximately 5 min and 36 longer videos ranging from 0.5 to 3 h, with a resolution of 1920 × 1080 pixels and a frame rate of 25 fps. To reduce computational load, improve efficiency, and ensure the capture of

essential information for recognition tasks, the following video processing steps were implemented. First, frames were extracted from all video clips using OpenCV, with a sampling rate of every 25 frames. This sampling strategy minimizes redundant frames and reduces computational demands. Next, the basic behaviors of cows were labeled using LabelImg. The labels included standing (stand), lying down (lie), eating (eat), and drinking (drink). After filtering and summarization, a total of 3458 images were labeled. The dataset was randomly split in an 8:1:1 ratio to create a training set (2766 images), a validation set (346 images), and a test set (346 images) for model training and evaluation. An example of cow behavior is illustrated in Figure 3. Figure 3a depicts the multi-scale behavior of cows in various scenes and complex backgrounds, including behaviors at both long and short distances. Figure 3b presents four examples of cow behavior: standing, drinking, eating, and lying down. Four behaviors relevant to cow health evaluation were selected as research subjects: standing, lying down, drinking, and eating. The criteria for determining cow behavior are presented in Table 2 [34].
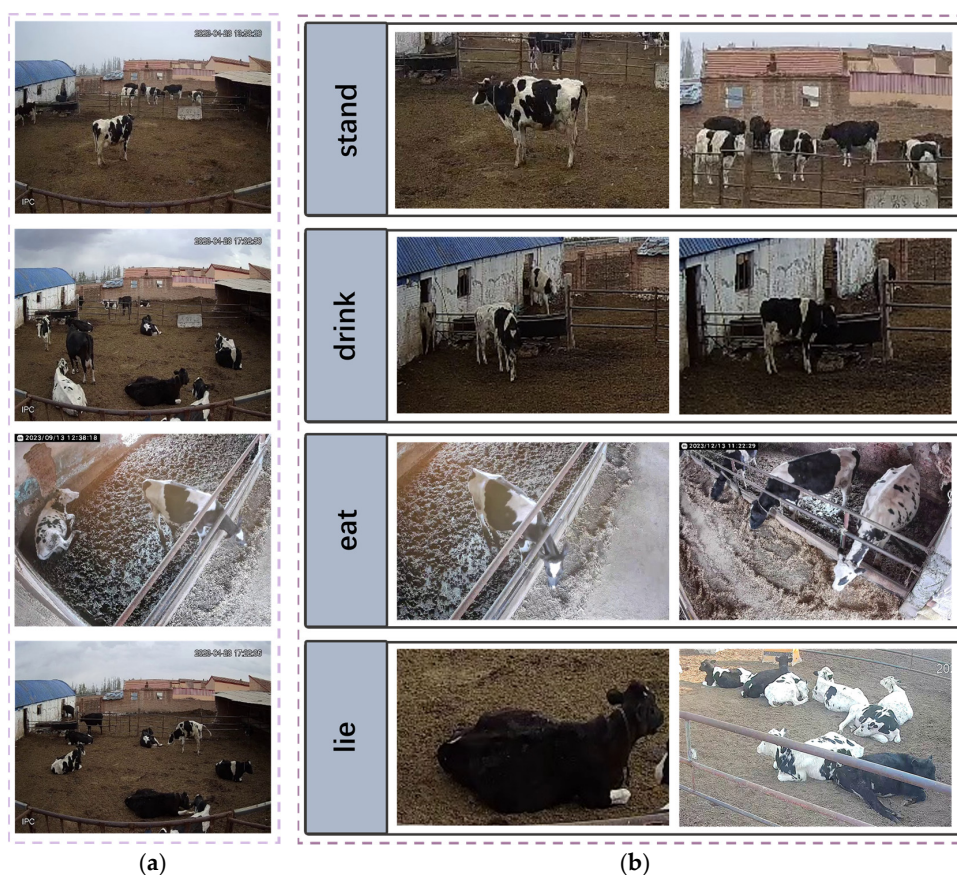

(**a**)   (**b**)

Figure 3. Example of a cow behavior shot. (**a**) Surveillance video scene; (**b**) example of cow behavior.

Table 2. Criteria for determining cow behavior.

Category	Judgment Standard	Visual Feature	Labels
Stand	Limbs upright, supporting body weight; abdomen off the ground.	Legs visible; body straight or near-straight; no ground contact.	stand
Lie	Belly or body touching the ground; limbs bent, relaxed posture.	Body close to the ground; outline horizontal; limbs bent.	lie
Eat	Head passes over railing, in contact with or near the feed.	Head near feed; body tilted; neck passing through railing.	eat
Drink	Standing upright; head positioned above sink, mouth touching water.	Head extended forward, contacting water source.	drink

Table 2 outlines four cow behaviors and their corresponding data. When standing, the cow has all four legs upright, supporting its weight, with the belly kept at a distance from the ground, the body in a vertical position, and the head either horizontal or lowered. When lying down, the belly, chest, or part of the body touches the ground, the limbs are bent, the posture is relaxed, the head is placed flat or slightly raised, and the cow lies horizontally. When eating, the limbs are upright, the head is lowered, passing through the railing, and the mouth touches or is close to the feed. When drinking, the limbs are upright, supporting the weight, the head is positioned above the water tank, and the mouth touches the water surface. Based on the definitions in Table 2, the dataset was labeled, resulting in 14,754 behaviors, including standing (6845), lying down (4139), eating (2280), and drinking (1490). Figure 4 presents a comparison of the tag statistics. Cows were most frequently observed standing, while eating and drinking were the least frequent behaviors. This distribution closely mirrors that observed in real farm environments and meets the requirements for cow behavior analysis in complex settings.
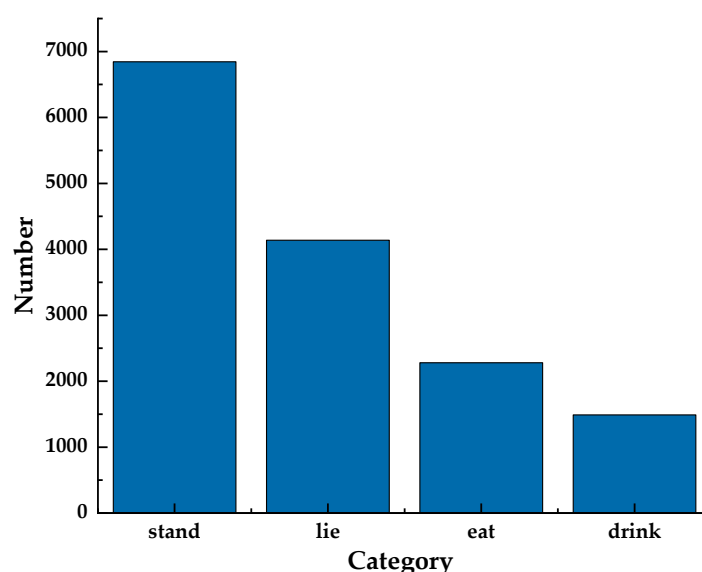

Figure 4. Behavioral labeling analysis of cows.

2.2.2. Training Platform

The computing platform used in this study consisted of a 12th Gen Intel® Core™ i7-12700H CPU (Intel Corporation, Santa Clara, CA, USA) processor with a clock speed of 2.30 GHz, 24 GB of RAM, 1 TB of SSD storage, and an NVIDIA GeForce RTX 3070 graphics card with 16 GB of VRAM. The software environment included a Windows operating system with CUDA 12.1 and Python 3.9 installed. Additionally, the deep learning framework PyTorch 2.1.0 was employed in this experiment. The hardware specifications are detailed in Table 3. All comparison algorithms were run in the same environment. The network training parameter settings are shown in Table 3 according to the test conditions.

Table 3. Training parameter setting.

Parameter	Values
Training batch size	16
Epochs	200
Image_size	640×640
Batch_size	8
Initial learning rate	0.01
Momentum	0.937

2.2.3. Model Evaluation Index

To verify the performance of this improved model, this study uses Precision (P), Recall (R), Average Precision (AP), mean Average Precision (mAP), Params, Floating Point operations (FLOPs), and F1 Score. F1 Score is an indicator used to evaluate the performance of a classification model, combining Precision and Recall. When the weight factor is $\alpha = 0.5$, compared to the commonly used F1 Score, P is more important than R and has twice the weight of R. mAP50 refers to the AP calculated for all images in each category when the threshold is set to 0.5, and then averaged across all categories. mAP50-95 refers to the AP calculated for all images in each category when the threshold is set to 0.5–0.95, and then averaged across all categories. The calculations of P, R, and mAP are shown in Equations (1)–(3) [35].

$$P = \frac{TP}{(TP + FP)} \times 100\% \tag{1}$$

$$R = \frac{TP}{(TP + FN)} \times 100\% \tag{2}$$

$$mAP = \frac{\sum_{i=1}^{C} AP(C)}{C} \times 100\% \tag{3}$$

Here, TP (True Positives) represents the number of correctly classified positive examples, FP (False Positives) denotes the number of negative examples incorrectly classified as positive, FN (False Negatives) indicates the number of positive examples incorrectly classified as negative, and C represents the number of detection categories, which in this study is C = 1.

2.3. A Cow Behavior Recognition Model Based on an Improved YOLOv5 Network

In this study, the YOLOv5 network was chosen as the foundational model for cow behavior recognition. To address the aforementioned challenges, the trained YOLOv5 model was enhanced to improve detection performance without compromising accuracy. Three distinct attention mechanisms were incorporated to enhance the model's sensitivity to small-scale cow behaviors, maintain effective detection of large-scale cows, and distinguish between behavioral similarities, such as posture. Without increasing network complexity or compromising real-time performance, four attention modules were integrated into the neck network of the YOLOv5 model. These modules were implemented in the upsampling and downsampling processes to enhance deep feature extraction of small-scale cow targets and focus on relevant feature information. Additionally, the C3 module was enhanced, and deformable convolution modules were incorporated to further improve the model's accuracy in identifying cow behavior characteristics. Finally, to optimize the detection head's performance, the original detection head was replaced with a dynamic head, enabling efficient and accurate detection of cow behavior across different scales in farm environments. The enhanced network structure is illustrated in Figure 5, with the modified modules highlighted in red. These modifications significantly enhanced the model's performance in detecting cow behavior in farm environments, particularly its ability to distinguish between cows of varying sizes exhibiting similar behaviors.

2.3.1. Add Shuffle Attention

Shuffle Attention (SA) is a hybrid attention mechanism that integrates channel attention and spatial attention mechanisms. It groups input feature maps and independently performs convolution operations within each group. It facilitates feature exchange and fusion between groups through channel shuffling operations and employs a dual attention strategy to enhance feature diversity and expressiveness, enabling more accurate capture of image details and contextual information [36]. Given the rich and diverse behaviors

of cows, each behavior exhibits distinct characteristics in the image. SA enhances the model's sensitivity to small-scale cow behaviors, improves feature expression richness and accuracy, and enables more precise capture of key behavioral features, thereby boosting the performance of multi-scale cow behavior recognition in complex backgrounds. The structure of the SA module is illustrated in Figure 6.
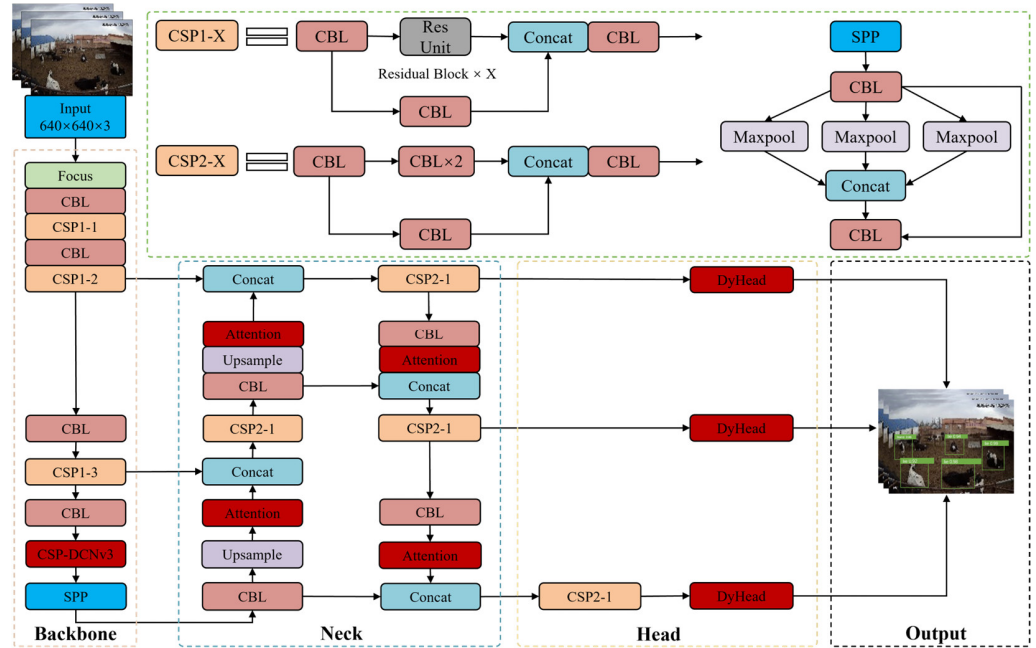

Figure 5. Improved network structure of YOLOv5 model.
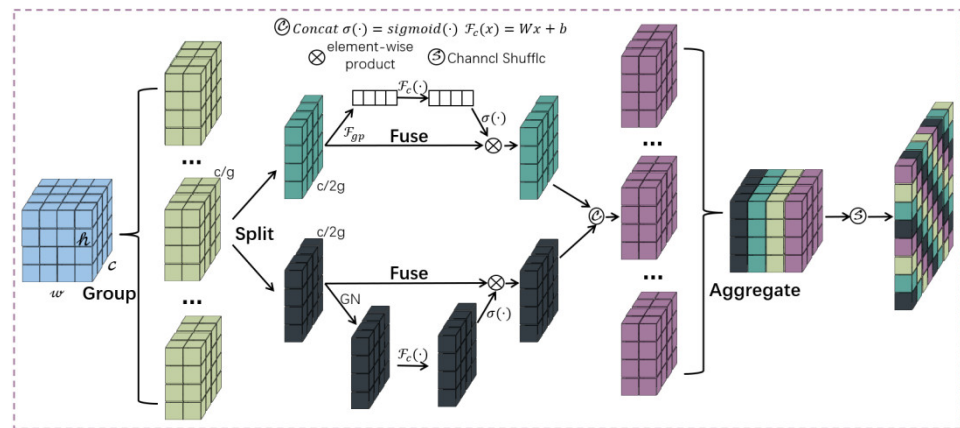

Figure 6. Shuffle Attention module structure.

2.3.2. Introduce Deformable Convolution DCNv3

DCNv3 integrates a sparse attention mechanism with convolution to process each output position in a sliding window manner. It dynamically samples points with an adaptive range and aggregates spatial features using input-dependent attention weights. This enables the model to focus more effectively on cows and relevant feature areas in complex backgrounds, adaptively adjust the receptive field based on changes in background, cow position, and posture, and improve feature extraction and robustness in complex environments. The principle of deformable convolution is illustrated in Figure 7.
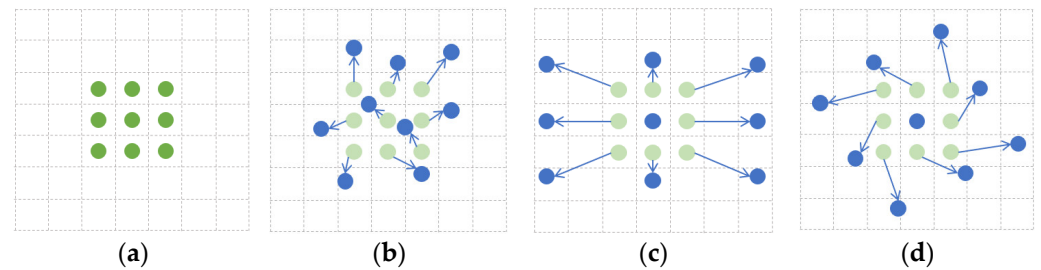
Figure 7. Principle of transformable convolution. (**a**) Conventional convolution; (**b**) deformable convolution; (**c**) special deformable convolution; (**d**) special deformable convolution. Green indicates the sampling points of the regular convolution, and blue indicates the dynamically sampled points of the deformable convolution.

This study focuses on detecting dairy cow behavior in farm environments. A third-generation deformable convolutional network (DCNv3) was integrated to enhance the detection performance of diverse cow behaviors. The implementation process is illustrated in Figure 8. Compared to traditional deformable convolution, DCNv3 not only captures target feature information but also minimizes the interference of irrelevant background information during the classification of cows with diverse behaviors. Significant performance improvements are achieved through the following key technical enhancements:

(1) Weight Sharing Mechanism: A weight-sharing mechanism reduces model complexity by splitting weights into two parts along the depth dimension. This mechanism simplifies the model structure and enhances processing efficiency. Specifically, weights along the depth dimension are regulated by the perception mechanism at the original position, while weights are shared across sampling points. This design reduces parameter count and optimizes the computational process.

(2) Multiple Group Segmentation Strategy: A grouping mechanism is implemented during spatial information aggregation. By dividing the sampling process into N groups, each with its own sampling offset and adjustment scale, the model achieves diverse spatial information aggregation modes within a single layer. This strategy enables the model to capture richer and more detailed feature information, enhancing the detection accuracy of diverse cow behaviors.

(3) SoftMax Normalization: To enhance training stability, the SoftMax function replaces the traditional Sigmoid function for normalization. This improvement addresses the gradient vanishing problem associated with Sigmoid, ensuring stable model training.
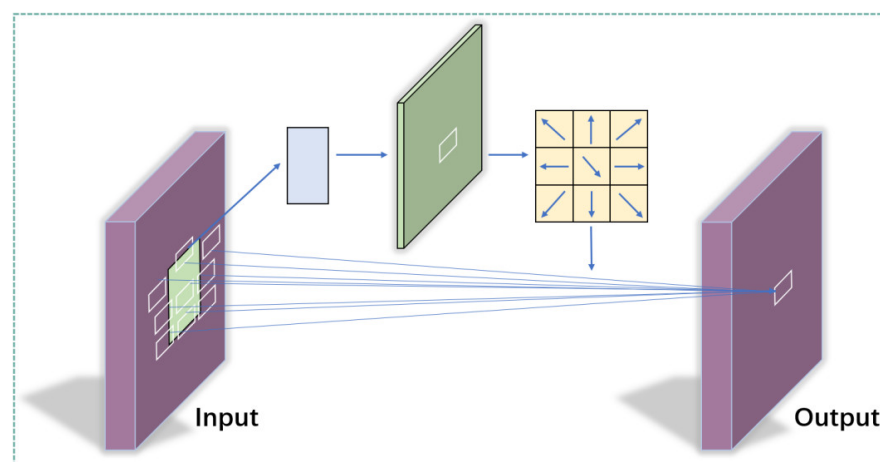
Figure 8. DCNv3 convolutional implementation process.

2.3.3. Replace Dynamic Head

Although traditional algorithms aim to enhance the detection head, they lack a unified approach. DyHead is an advanced object detection head framework specifically designed to address the challenges of complex scenes, such as detecting dairy cow behavior in farm environments. These scenes are characterized by varying object sizes, a wide range of behaviors (e.g., feeding, lying down, standing, and drinking), and similarities between certain behaviors. DyHead effectively adapts to this diversity and complexity through a dynamic attention mechanism. Its structure is illustrated in Figure 9. It employs scale-aware, spatial-aware, and task-aware attention modules to manage detection tasks for diverse objects in a unified and flexible manner, enabling the algorithm to better understand and recognize various cow behaviors.
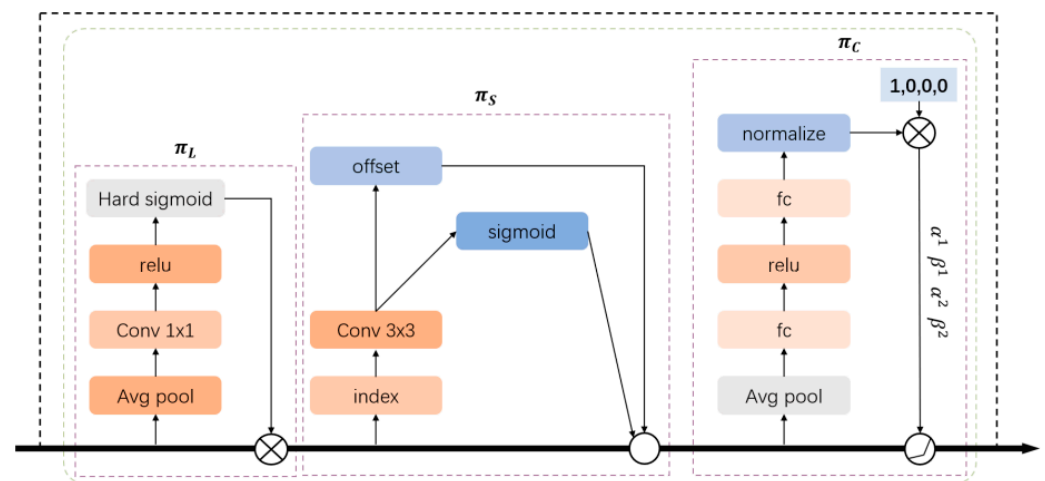
Figure 9. DyHead structure.

3. Results

In this study, model performance was evaluated by monitoring the loss rate and accuracy during training, with frozen layers and transfer learning techniques employed to leverage prior knowledge and accelerate learning. The training process comprised 200 epochs, with a partial layer freezing strategy applied during the first 50 epochs. This strategy aims to mitigate drastic adjustments to pre-trained parameters, ensuring steady optimization without significant disruptions. Trends in the loss function and mAP indicate that the model loss gradually decreases over time, while mAP steadily increases. The variation in the loss function during training is illustrated in Figure 10. Although the initial loss was high, the error rate decreased significantly as training progressed, and detection accuracy improved gradually. The variation in model accuracy during training is depicted in Figure 11. This phenomenon not only validates the effectiveness of the early freezing strategy but also demonstrates its ability to prevent overfitting and convergence issues in the early training phase. As the training process advanced, the model became more effective at learning and identifying key features for detecting dairy cow behavior on the farm. Overall, the training strategy employed in this study improved learning efficiency, enhanced the model's generalization ability, and increased overall detection accuracy.
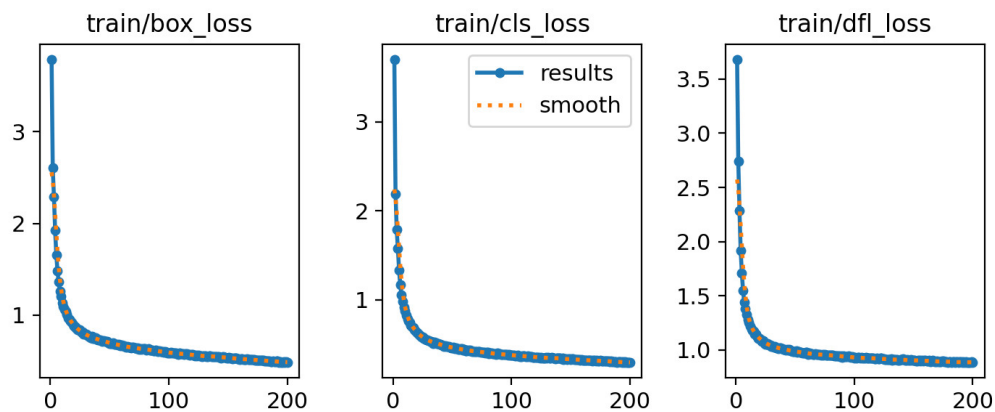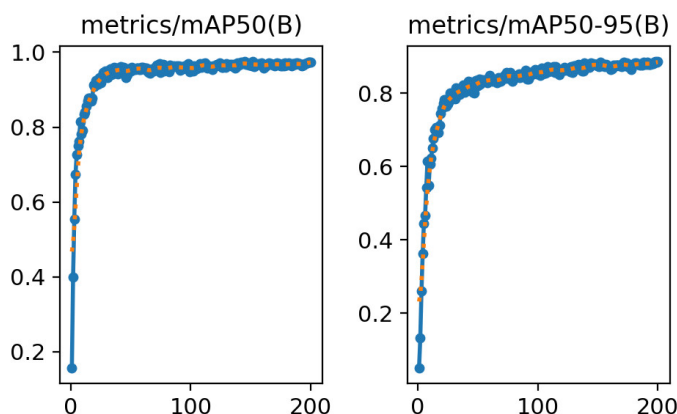
Figure 10. Loss function varies during training.

Figure 11. Model accuracy changes during training.

3.1. Comparison of the Results of the Cow Behavior Recognition Model

3.1.1. Comparison of the Performance of Different Models

In order to evaluate the performance of different models in cow behavior recognition, this paper conducts a comprehensive analysis of indicators such as Params, FLOPs, Precision, Recall, F1 Score, and mAP by training and testing the models on a self-built cow behavior dataset. The experimental results are presented in Table 4. The experimental results clearly demonstrate the advantages and disadvantages of each model in recognizing cow behavior in complex scenarios.

Table 4. Comparison of performance metrics of different models.

Index Algorithms	Params (M)	FLOPs (G)	Precision (%)	Recall (%)	F1 Score (%)	mAP (%)
Faster R-CNN	27.4	74.0	79.2	78.3	78.7	78.7
YOLOv3	7.3	20.6	76.0	77.3	76.4	79.0
YOLOv4	4.6	10.7	89.3	89.3	89.3	91.5
YOLOv5	1.8	4.8	93.1	90.0	91.1	94.0
YOLOv8	2.3	5.7	90.3	91.4	89.5	92.6
YOLOx	12.1	52.78	87.5	89.9	87.7	89.2
YOLO11	2.6	6.3	92.5	91.3	91.7	93.2
Improved YOLOv5	3.2	6.8	93.8	96.7	95.2	97.7

The improved YOLOv5 model shows significant advantages in all performance indicators, especially in terms of mean Average Precision (mAP, 97.7%) and Recall (Recall, 96.7%), which are significantly higher than those of other comparison models. Specifically,

the SA module significantly enhances the model's feature extraction ability for small-scale targets by fusing channel attention and spatial attention mechanisms. In particular, it can effectively suppress background noise interference and improve the model's ability to focus on key behavior regions in the task of recognizing cow behavior in complex backgrounds. The DCNv3 module enhances the model's adaptability to changes in cow posture and complex shapes by dynamically adjusting the receptive fields of the convolutional kernels, enhancing the model's adaptability to changes in cow posture and complex shapes, thereby exhibiting greater robustness in multi-scale object detection tasks. The DyHead module, on the other hand, optimizes the model's detection capabilities for objects of different scales by dynamically assigning feature weights, significantly improving the mAP metric.

The improved YOLOv5 model also excels in terms of computational efficiency. Although its Params (3.2 M) and FLOPs (6.8 G) are slightly higher than those of the original YOLOv5 model (1.8 M, 4.8 G), they are much lower than those of Faster R -CNN (27.4 M, 74.0 G) and YOLOx (12.1 M, 52.78 G), indicating that the improved model has higher computational efficiency and deploy ability while maintaining high recognition accuracy.

In addition, the improved model performs particularly well in complex background and multi-scale object detection tasks. In a scene with a large number of cows, the improved model can effectively deal with object occlusions and pose changes, significantly reducing the false detection rate and missed detection rate. In contrast, although YOLOv8 and YOLO11 perform better in terms of parameters and computing power, their mAP metrics are 5.1 and 4.5 percentage points lower than those of the improved model, respectively, indicating that there is still much room for improvement in their adaptability in complex scenes.

3.1.2. Comparison of the Results of Identifying Different Types of Targets

The results of the recognition of different cow behaviors reflect the model's ability to extract and distinguish target features. Therefore, the recognition results (AP values) of the eight models for a single category are compared, as shown in Table 5. The improved YOLOv5 model shows significant advantages in all behavior categories, especially in the standing (98.1%) and lying (98.9%) behaviors. The AP value for the standing behavior (Stand) reached 98.1%, an increase of 2.9 percentage points compared to the original YOLOv5 model's 95.2% by 2.9 percentage points; the AP value for lying behavior (Lie) reached 98.9%, an increase of 4.4 percentage points over the original YOLOv5 model's 94.5%. This is because the standing and lying behaviors are relatively stable in posture and therefore easier to detect. In addition, in the complex farm environment, cows may have similar postures but different scales. SA combines channel attention and spatial attention mechanisms to enhance the model's ability to focus on small-scale targets and suppress irrelevant background noise, which helps improve the ability to recognize these behaviors. The model's AP values for eating behavior (Eat) and drinking behavior (Drink) were 97.7% and 96.1%, respectively, which is an improvement of 5.2% and 2.3% over the original YOLOv5 model. This is because drinking and eating involve more dynamic postures and specific equipment, such as obvious licking movements of the head through the railing when eating, and standing near the drinking trough with the head positioned above it when drinking. The DCNv3 module improves the model's ability to adapt to these changes by dynamically adjusting the receptive field, while the dynamic head detection DyHead ensures accurate detection at different scales.

Table 5. Comparison of target recognition Average Precision values of different categories.

Model	Params (M)	FLOPs (G)	Average Precision AP(%)				mAP (%)
			Stand	**Lie**	**Eat**	**Drink**	
Faster R-CNN	27.4	74.0	78.7	81.4	77.6	77.1	78.7
YOLOv3	7.3	20.6	80.1	78.4	79.4	78.1	79.0
YOLOv4	4.6	10.7	91.5	93.0	90.3	91.2	91.5
YOLOv5	1.8	4.8	95.2	94.5	92.5	93.8	94.0
YOLOv8	2.3	5.7	93.7	92.6	91.3	92.7	92.6
YOLOx	12.1	52.78	88.6	89.4	91.3	87.6	89.2
YOLO11	2.6	6.3	92.8	97.1	91.2	91.8	93.2
Improved YOLOv5	3.2	6.8	98.1	98.9	97.7	96.1	97.7

Compared with other models, the improved YOLOv5 model performed particularly well in standing and lying behaviors. Faster R-CNN's AP value for standing behavior was 78.7%, 19.4 percentage points lower than the improved YOLOv5 model; YOLOv3's AP value for lying behavior was 78.4%, 20.5 percentage points lower than the improved YOLOv5 model. YOLOv4 performed better in standing and lying behavior, its AP values were 91.5% and 93.0%, respectively, which were still significantly lower than those of the improved YOLOv5 model. YOLOv8 and YOLO11 had AP values of 93.7% and 92.6%, and 92.8% and 97.1% for standing and lying behavior, respectively, which were better, but still not as good as the improved YOLOv5 model. The improved YOLOv5 model also performed well in the recognition of dynamic behaviors such as eating and drinking. For example, the improved YOLOv5 model achieved an AP value of 97.7% for eating behavior, which is an improvement of 6.4% over YOLOv8's 91.3% and 6.5% over YOLO11's 91.2%. For drinking behavior, the AP value of the improved YOLOv5 model was 96.1%, which was 3.4% and 4.3% higher than YOLOv8's 92.7% and YOLO11's 91.8%, respectively. These results further verify the applicability and reliability of the improved YOLOv5 model in real farm environments.

3.2. Ablation Experiment

3.2.1. Performance Analysis of DCNv3 and DyHead

The analysis of Average Precision (AP) values across all behavior categories demonstrates the effectiveness of the improved YOLOv5 model. To assess the impact of DCNv3 and DyHead on the recognition of different behavior categories, ablation experiments based on YOLOv5 were conducted, and the results are presented in Table 6.

Table 6. Analysis of ablation experiments with DCNv3 and DyHead models.

Index Algorithms	Precision (%)	Recall (%)	F1 Score (%)	mAP (%)
YOLOv5	93.1	90.0	91.5	94.0
YOLOv5 + DCNv3	94.1	94.5	94.3	95.8
YOLOv5 + DyHead	95.2	92.6	93.9	96.3
YOLOv5 + DCNv3 + DyHead	94.5	95.1	94.9	96.6

Table 6 illustrates the contributions of DCNv3 and DyHead to enhancing the performance of YOLOv5. The ablation experiment results indicate that these two modules significantly improve the accuracy and robustness of cow behavior recognition. Specifically, with the introduction of DCNv3 alone, the model's Precision increased from 93.1% to

94.1%, Recall from 90.0% to 94.5%, F1 Score from 91.5% to 94.3%, and mean Average Precision (mAP) to 95.8%, highlighting DCNv3's significant advantages in enhancing feature extraction, adapting to complex backgrounds, and handling multi-scale targets. On the other hand, the addition of DyHead improved Precision and Recall to 95.2% and 92.6%, respectively, while F1 Score and mAP reached 93.9% and 96.3%, confirming DyHead's effectiveness in dynamic target detection and feature fusion. When DCNv3 and DyHead are combined, model performance is further optimized, with Precision, Recall, F1 Score, and mAP reaching 94.5%, 95.1%, 94.9%, and 96.6%, respectively. This result demonstrates that DCNv3 and DyHead exhibit strong synergy in capturing key features, optimizing object boundaries, dynamically assigning feature weights, and enhancing model robustness, providing reliable technical support for cow behavior detection in complex scenes.

3.2.2. The Influence of Different Attention Mechanisms on the Performance of the Improved Model

The experimental results indicate that incorporating an attention mechanism further optimizes the model's recognition performance, with SA achieving the best results. Without an attention mechanism, the model achieved Precision, Recall, F1 Score, and mean Average Precision (mAP) values of 94.5%, 95.1%, 94.9%, and 96.6%, respectively. With the addition of the Squeeze-and-Excitation (SE) attention mechanism, Precision increased to 96.4%, but Recall decreased to 90.7%, indicating its strengths in accurate object detection but limitations in object recall. In contrast, the Efficient Multi-Scale Attention (EMA) mechanism improved Recall to 92.5% while maintaining a high Precision of 95.4%, demonstrating its balanced feature expression. However, with the introduction of SA, the model's performance reached its peak, with Precision, Recall, F1 Score, and mAP increasing to 93.8%, 96.7%, 95.2%, and 97.7%, respectively. This significant improvement demonstrates that SA effectively enhances the model's target recognition ability in complex scenes by strengthening the fusion of channel and spatial features. In particular, SA exhibits excellent robustness in handling multi-scale targets and complex behavioral features, providing critical technical support for further model optimization. Table 7 presents the performance metrics of the improved model with different attention mechanisms.

Table 7. Performance metrics of different attention mechanisms for improved models.

Index Algorithms	Precision (%)	Recall (%)	F1 Score (%)	mAP (%)
YOLOv5	93.1	90.0	91.5	94.0
Iattentive + DCNv3 + DyHead	94.5	95.1	94.8	96.6
SE + DCNv3 + DyHead	96.4	90.7	93.4	95.7
EMA + DCNv3 + DyHead	95.4	92.5	93.9	97.3
SA + DCNv3 + DyHead	93.8	96.7	95.2	97.7

3.3. Cows Behavior Recognition Results Based on Improved YOLOv5 Network

To evaluate the performance of the improved model in real-world scenarios, 480 images not included in the training set were selected for testing. These images encompass a wide range of cow behaviors across various farm environments, ensuring a comprehensive and unbiased evaluation. The detection results are illustrated in Figure 12. The results demonstrate that the improved model achieves high accuracy in identifying diverse cow behaviors in farm environments. Even under complex backgrounds or suboptimal camera angles, the model accurately detects cow behaviors, highlighting its reliability and adaptability in practical applications.
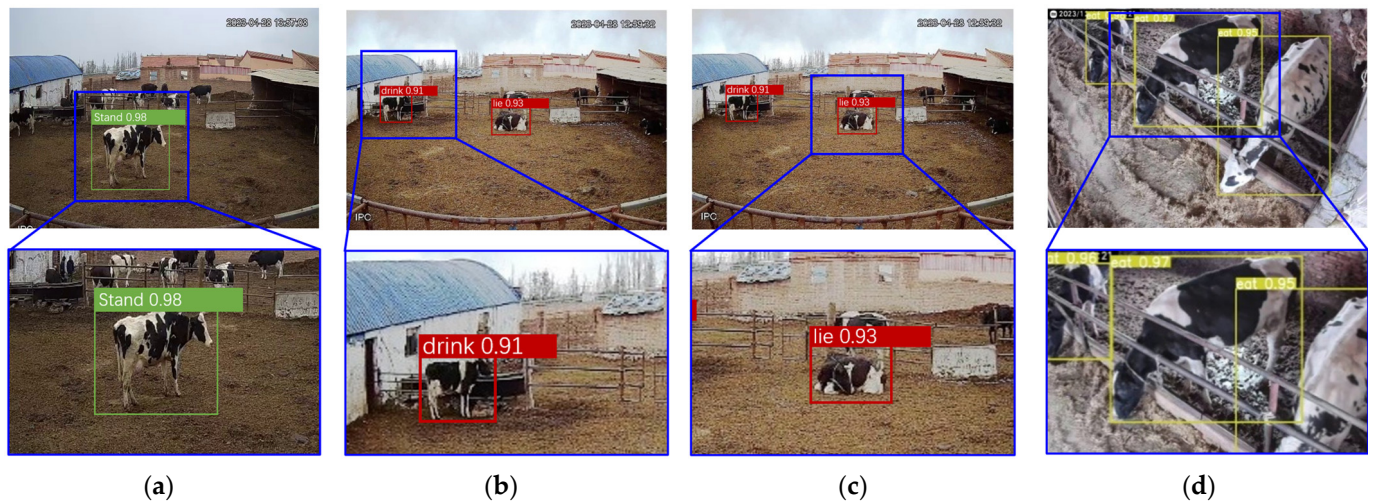
Figure 12. Improved recognition effect of the YOLOv5 model. (**a**) Standing behavior of outdoor dairy cows; (**b**) drinking behavior of outdoor dairy cows; (**c**) lying behavior of outdoor dairy cows; (**d**) feeding behavior of indoor dairy cows.

4. Discussion

4.1. Analysis of Missed and Misdiagnosed Cow Behaviors

To better understand the role of attention mechanisms in detecting dairy cow behavior on farms, Gradient-weighted Class Activation Mapping (Grad-CAM) technology was employed to visualize the model's focus areas under different attention mechanisms. Grad-CAM is a visualization tool for deep learning models that highlights key image regions in the model's decision-making process, aiding in the analysis of model performance in behavior detection under complex backgrounds. Figure 13 illustrates the visualization results of behavior detection tasks for models without attention mechanisms and those with different attention mechanisms.

Comparison of the results reveals that attention mechanisms significantly enhance the model's ability to focus on key cow behavior areas, particularly in complex or noisy backgrounds. As shown in the figure, the model without an attention mechanism exhibits a scattered focus when handling multiple target occlusions and ambiguous behavior regions, resulting in missed and false detections. However, the Squeeze-and-Excitation (SE) and Efficient Multi-Scale Attention (EMA) models improve feature extraction to some extent, enhancing focus on key behavior regions. SE exhibits attention deviation in some complex scenes, such as misidentifying background areas as targets, leading to detection errors. EMA, on the other hand, shows minor omissions in detecting dynamic behaviors (e.g., eating and drinking), likely due to insufficient adaptability to dynamic behavioral changes.

In contrast, the SA hybrid model demonstrated the best performance across all behavior categories. Its Grad-CAM images displayed precise attention to key feature areas, reducing reliance on background interference and more accurately capturing key cow behavior features. This performance enabled SA to achieve higher detection accuracy and stability in multi-objective complex environments, effectively reducing missed and misdiagnosed cases and providing more reliable technical support for cow behavior detection.
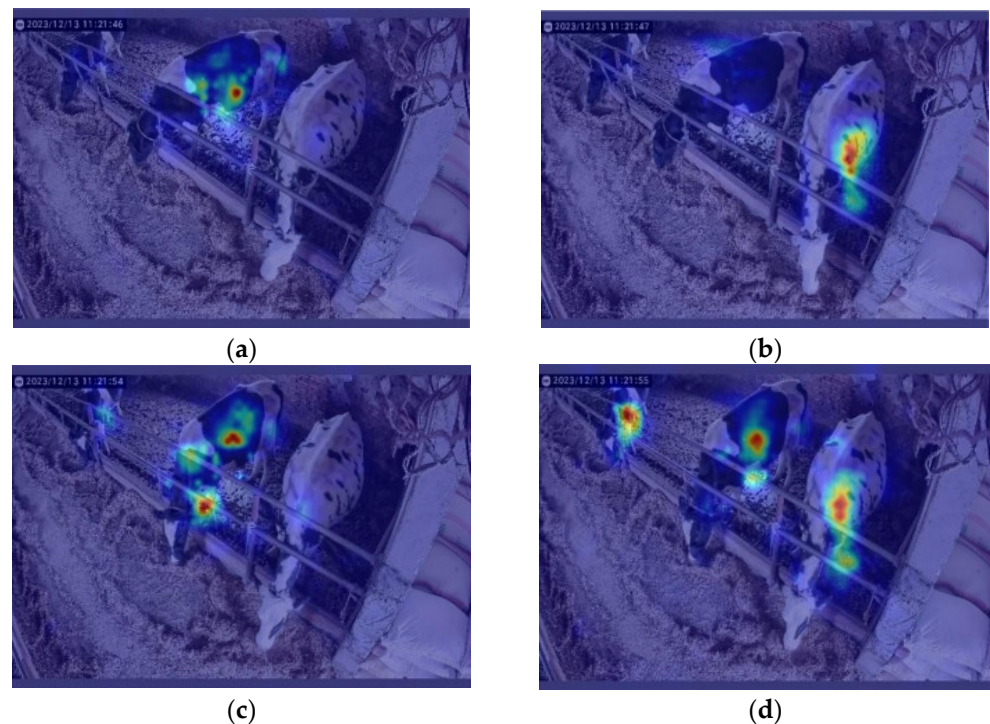
Figure 13. Visual comparison of original model and attention model. (**a**) No attention mechanism; (**b**) SE attention mechanism; (**c**) EMA attention mechanism; (**d**) SA attention mechanism. Red indicates the highest attention weight; yellow indicates a medium-high attention weight; green indicates a medium attention weight; blue indicates a low attention weight; dark blue indicates the lowest attention weight.

4.2. Comparison of the Improved Model with the Results of Previous Studies

This study integrates four SA mechanisms into YOLOv5 to enhance small-scale target feature extraction, improves multi-behavior recognition accuracy through DCNv3 in the C3 module, and replaces the original detection head with DyHead to improve recognition efficiency and accuracy in complex environments and for multi-scale targets. Table 8 summarizes key research findings on cow behavior recognition in recent years and compares them with this study.

This study demonstrates clear advantages in several aspects. In terms of behavior category coverage, this study includes four behaviors: standing, drinking, eating, and lying down, whereas most existing studies focus on single-behavior detection (e.g., eating or estrus climbing behavior), limiting their applicability to multi-behavior monitoring in real farm settings. In terms of data scale, this study's dataset comprises 1044 video segments, significantly larger than the 406 segments in the RexNet 3D-based study, greatly enhancing the model's generalization ability. Additionally, the mean Average Precision (mAP) of this study reached 97.7%, demonstrating excellent recognition performance in complex backgrounds and multi-scale target scenarios. However, the mAP of the YOLOv3-based cow behavior recognition study was only 83.8%, which cannot meet the requirements of complex real-world scenarios.

In terms of model structure optimization, this study incorporates SA, DCNv3, and DyHead, enhancing the model's detection capabilities for complex backgrounds and multi-scale targets while maintaining real-time performance. This study is superior to the study based on EfficientNet-LSTM in terms of accuracy. Overall, this study not only enhances the detection accuracy of multiple behavior categories but also demonstrates higher adaptability in complex environments, offering an effective solution for real-time cow behavior monitoring.

Table 8. Comparison of improved model with results of previous studies.

Year	Model	Behaviors	Dataset	Sampling Rate (fps)	Environment	Cameras	mAP (%)
2016	Clustering	Lie, stand, walk, run, jump	162 videos (-)	25	Indoor	1	97.3
2018	KNN	Limp	360 videos (30 cows)	25	Outdoor	1	82.7
2019	CNN	Estrus (mounting)	25,000 videos (50 cows)	30	Outdoor	2	98.2
2020	YOLOv3	Eat	1846 images	24	Indoor	1	83.8
2020	EfficientNet-LSTM	Lie, stand, walk, drink, feed	1009 videos (-)	10	Outdoor	1	97.8
2021	YOLOv3	Estrus (mounting)	3600 videos (56 cows)	5	Outdoor	2	98.1
2022	RexNet 3D	Lie, stand, walk	10 videos (30 cows)	5	Outdoor	2	95.0
2024	Improved YOLOv5	Lie, stand, eat, drink	1044 videos (38 cows)	25	Indoor + Outdoor	2	97.7

5. Conclusions

5.1. Summary

(1) The proposed improved model effectively addresses the challenges of complex backgrounds and multi-scale targets by integrating three modules: SA, DyHead, and DCNv3. Specifically, SA significantly enhances the model's focus on behavioral regions by combining channel attention and spatial attention mechanisms, while suppressing irrelevant features in complex backgrounds. DyHead optimizes multi-scale target detection by dynamically adjusting feature weights, particularly enhancing the ability to distinguish between small targets and those with similar behavioral features. DCNv3 improves the model's robustness in occluded scenes and complex shapes by dynamically adjusting the receptive field and adaptively sampling feature regions.

(2) The results indicate that the model incorporating both DCNv3 and DyHead performs best, achieving a 2.6% increase in mAP50 and a 1.8% increase in mAP50-95 compared to the original model. The model integrates DCNv3, DyHead, and an attention mechanism to address these challenges, improving accuracy and efficiency by assigning different weights to different input components. The results demonstrate that the improved model, combining SA, DCNv3, and DyHead, significantly enhances multi-scale cow behavior recognition, achieving a 3.7% increase in mAP over the original model. The recognition results indicate that the improved model excels at distinguishing and identifying cow behaviors, surpassing the accuracy of the YOLOv5 model. In tests using images from natural environments, the improved model demonstrated high-precision cow behavior recognition. This advantage significantly enhances the model's ability to recognize multiple cow behaviors in real-world scenarios.

5.2. Prospect

(1) The current focus is on optimizing model structure. Future work can further enhance model performance on edge computing devices and achieve higher execution efficiency by integrating advanced techniques such as pruning.

(2) This study focuses on four basic cow behaviors: standing, lying, drinking, and eating. To comprehensively assess cow health and reproductive status, future research should

expand to accurately identify complex behaviors such as ruminating, tail-wagging, and mounting, enabling more precise health predictions.

References

1. Cheng, G.; Wu, J.; Xing, L.; Zhu, M.; Zhang, J.; Han, S. Fine-grained cows' behavior classification method based on IMU. *J. China Agric. Univ.* **2022**, *27*, 179–186.
2. He, D.; Liu, D.; Zhao, K. Review of Perceiving Animal Information and Behavior in Precision Livestock Farming. *Trans. Chin. Soc. Agric. Mach.* **2016**, *47*, 231–244.
3. Wang, Z.; Song, H.; Wang, Y.; Hua, Z.; Li, R.; Xu, X. Research progress and technology trend of intelligent morning of dairy cow motion behavior. *Smart Agric.* **2022**, *4*, 36–52.
4. Wang, X.; Siqinbate; Turigengbaiyila. TURIGEN Baiyila.Research Progress of Dairy Cattle Lying Behavior. *Mod. Anim. Husb.* **2017**, *5*, 3–5. [CrossRef]
5. Bian, S. Factors affecting the duration of lying rest in dairy cows. *Chin. Dairy Cow* **2017**, *7*, 6–9. [CrossRef]
6. Teng, G. Information sensing and environment control of precision facility livestock and poultry farming. *Smart Agric.* **2019**, *1*, 3.
7. Liu, J.; Jiang, B.; He, D.; Song, H. Individual Recognition of Dairy Cattle Based on Gaussian Mixture Model and CNN. *Comput. Appl. Softw.* **2018**, *35*, 159–164.
8. Yongliang, Q.; He, K.; Cameron, C.; Sabrina, L.; Daobilige, S.; Stuart, E.; Salah, S. Intelligent Perception-Based Cattle Lameness Detection and Behaviour Recognition: A Review. *Animals* **2021**, *11*, 3033. [CrossRef] [PubMed]
9. Chu, M.; Liu, X.; Zeng, X.; Wang, Y.; Liu, G. Research advances in the automatic detection technology for mastitis of dairy cows. *Trans. Chin. Soc. Agric. Eng.* **2023**, *39*, 1–12.
10. Zhao, K.; Li, G.; He, D. Fine Segment Method of Cows'Body Parts in Depth Images Based on Machine Learning. *Trans. Chin. Soc. Agric. Mach.* **2017**, *48*, 173–179.
11. Stygar, A.H.; Gómez, Y.; Berteselli, G.V.; Costa, E.D.; Canali, E.; Niemi, J.K.; Llonch, P.; Pastell, M. A Systematic Review on Commercially Available and Validated Sensor Technologies for Welfare Assessment of Dairy Cattle. *Front. Vet. Sci.* **2021**, *8*, 634338. [CrossRef] [PubMed]
12. Wang, Z.; Zong, Z.; Wang, H.; Yang, J.; Liu, G.; Du, Y. Research status and development of digital detection methods for cow estrus. *China Feed* **2021**, *1*, 134–138.
13. Smith, D.; Rahman, A.; Bishop-Hurley, G.J.; Hills, J.; Shahriar, S.; Henry, D.; Rawnsley, R. Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Comput. Electron. Agric.* **2016**, *131*, 40–50. [CrossRef]
14. Guo, Y.; He, D.; Chai, L. A Machine Vision-Based Method for Monitoring Scene-Interactive Behaviors of Dairy Calf. *Animals* **2020**, *10*, 190. [CrossRef]
15. Song, H.; Wu, D.; Yin, X.; Jiang, B.; He, D. Detection of cow breathing behavior based on the Lucas-Kanade sparse optical flow algorithm. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 215–224.
16. He, D.; Meng, F.; Zhao, K.; Zhang, Z. Basic behavior recognition of calves based on video analysis. *Trans. Chin. Soc. Agric. Mach.* **2016**, *47*, 294–300.

17. Song, H.; Jiang, B.; Wu, Q.; Li, T.; He, D. A cow lameness detection method based on fitting the slope feature of the straight line to the contour of the head and neck. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 190–199.

18. Zhang, S.; Tian, J.; Jian, L.; Ji, Z. A method for recognizing fighting behavior in piglets based on inter-frame difference method and single-point multi-frame detector. *J. Jiangsu Agric. Sci.* **2021**, *37*, 397–404.

19. Mao, Y.; Niu, T.; Wang, P.; Song, H.; He, D. Multi-objective cow mouth tracking and rumination monitoring using Kalman filtering and Hungarian algorithm. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 192–201.

20. Song, H.; Niu, M.; Ji, C.; Li, Z.; Zhu, Q. Multi-objective cow ruminant behavior monitoring based on video analysis. *Trans. Chin. Soc. Agric. Eng.* **2018**, *34*, 211–218.

21. Gu, J.; Wang, Z.; Gao, R.; Wu, H. A method for recognizing cow behavior based on fusing images and motion quantities. *Trans. Chin. Soc. Agric. Mach.* **2017**, *48*, 145–151.

22. Zhenwei, Y.; Yuehua, L.; Sufang, Y.; Ruixue, W.; Zhanhua, S.; Yinfa, Y.; Fade, L.; Zhonghua, W.; Fuyang, T. Automatic Detection Method of Dairy Cow Feeding Behaviour Based on YOLO Improved Model and Edge Computing. *Sensors* **2022**, *22*, 3271. [CrossRef] [PubMed]

23. Chen, C.; Zhu, W.; Norton, T. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Comput. Electron. Agric.* **2021**, *187*, 106255. [CrossRef]

24. Achour, B.; Belkadi, M.; Filali, I.; Laghrouche, M.; Lahdir, M. Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on Convolutional Neural Networks (CNN). *Biosyst. Eng.* **2020**, *198*, 31–49. [CrossRef]

25. Wang, Z.; Wang, S.; Wang, C.; Zhang, Y.; Zong, Z.; Wang, H.; Su, L.; Du, Y. A Non-Contact Cow Estrus Monitoring Method Based on the Thermal Infrared Images of Cows. *Agriculture* **2023**, *13*, 385. [CrossRef]

26. Zhang, H.; Wu, J.; Li, Y.; Li, S.; Wang, H.; Song, R. Research on the identification of the feeding behavior of multi-purpose beef cattle. *Trans. Chin. Soc. Agric. Mach.* **2020**, *51*, 259–267.

27. Liu, Z.; He, D. A method for recognizing the estrus behavior of cows based on convolutional neural networks. *Trans. Chin. Soc. Agric. Mach.* **2019**, *50*, 186–193.

28. Wang, S.; He, D. Research on the recognition of estrus behavior in dairy cows based on an improved YOLO v3 model. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 141–150.

29. Yin, X.; Wu, D.; Shang, Y.; Jiang, B. Huaibo SongUsing an EfficientNet-LSTM for the recognition of single Cow's motion behaviours in a complicated environment. *Comput. Electron. Agric.* **2020**, *177*, 105707. [CrossRef]

30. Ma, S.; Zhang, Q.; Li, T.; Song, H. Basic motion behavior recognition of single dairy cow based on improved Rexnet 3D network. *Comput. Electron. Agric.* **2022**, *194*, 106772. [CrossRef]

31. Zhang, Q.-L.; Yang, Y.-B. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. *arXiv* **2021**, arXiv:2102.00240.

32. Wang, G.; Li, H.; Ye, S.; Zhao, H.; Ding, H.; Xie, S. RFWNet: A Multiscale Remote Sensing Forest Wildfire Detection Network With Digital Twinning, Adaptive Spatial Aggregation, and Dynamic Sparse Features. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4708523. [CrossRef]

33. Luo, Z.; Wang, C.; Qi, Z.; Luo, C. LA_YOLOv8s: A lightweight-attention YOLOv8s for oil leakage detection in power transformers. *Alex. Eng. J.* **2024**, *92*, 82–91. [CrossRef]

34. Bai, Q.; Gao, R.; Zhao, C.; Li, Q.; Wang, R.; Li, S. Multi-scale behavior recognition method for dairy cows based on improved YOLOV5s network. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 163–172.

35. Wang, Z.; Xu, X.; Hua, Z.; Shang, Y.; Duan, Y.; Song, H. Lightweight recognition for the oestrus behavior of dairy cows combining YOLO v5n and channel pruning. *Trans. Chin. Soc. Agric. Eng.* **2022**, *38*, 130–140.

36. Jannu, C.; Vanambathina, S.D. Shuffle Attention U-Net for Speech Enhancement in Time Domain. *Int. J. Image Graph.* **2023**, *24*, 2450043. [CrossRef]