Research on Multi-target Cow Behavior Recognition Method Based on Deep Learning

Jizhen WU[1], **Jianfei SHI**[1*], **Zhiyuan JING**[2]

1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China; 2. Daqing Oilfield Design Institute Co., Ltd., Daqing 163001, China

Abstract To address the issue of low recognition accuracy for eight types of behaviors including standing, walking, drinking, lying, eating, mounting, fighting and limping in complex multi-cow farm environments, a multi-target cow behavior recognition method based on an improved YOLOv11n algorithm was proposed. The detection capability for small targets in images was enhanced by incorporating a DASI module into the backbone network and a MDCR module into the neck network, based on YOLOv11. The improved YOLOv11 algorithm increased the mean average precision from the original 89.5% to 93%, with particularly notable improvements of 8.7% and 6.3% in the average precision for recognizing drinking and walking behaviors, respectively. These results fully demonstrate that the proposed method enhances the model's ability to recognize cow behaviors.

Key words Image recognition; YOLOv11n; Cow behavior recognition; Deep learning
DOI:10.19759/j.cnki.2164 – 4993.2025.06.008

The dairy farming industry has very broad prospects. With advancements in science and technology and the growing consumer demand for high-quality dairy products, the precise and efficient recognition of daily cow behaviors is of great significance. Cow behavior can genuinely reflect their living conditions and serves as an important basis for assessing their health status. Specifically, mounting behavior in dairy cows can effectively reflect their estrus status, contributing to improved reproductive efficiency and offspring quality[1]. Abnormal behavioral patterns such as limping often indicate potential disease risks and require heightened attention[2].

The concept of image recognition can be traced back to the 1950s, when research primarily focused on pattern recognition and simple image processing techniques. In 1986, Rumelhart, Hinton and Williams[3] published the backpropagation algorithm, providing an effective method for training neural networks. The development of animal posture and behavior recognition has undergone a transition from traditional methods to the application of modern technologies. The emergence of digital image processing techniques has brought new possibilities for animal posture and behavior recognition. Through the analysis of animal images, features such as animal postural characteristics and movement trajectories can be extracted, enabling the automatic identification and analysis of animal behavior. Du[4] proposed a method for recognizing daily behaviors of dairy cows, which utilizes the YOLOv5s object detection algorithm to identify routine cow behaviors and includes a designed algorithm for calculating the duration of individual cow behaviors.

Received: August 13, 2025 Accepted: October 17, 2025
Supported by The Three Vertical Basic Cultivation Project of Heilongjiang Bayi Agricultural University (ZRCPY202314).
Jizhen WU (2000 –), male, P. R. China, master, devoted to research about livestock information.
* Corresponding author.

Dataset Construction

From September to October 2023, outdoor data were collected at an open grassland pasture in Chen Barag Banner, Hulunbeier City, Inner Mongolia Autonomous Region, along the S201 highway. Videos of cattle herd activities were recorded using a handheld iQOO Z8X smartphone with a 50-megapixel camera, at a resolution of 1 440 × 1 080 pixels, mounted on a fixed stand. The video format was MP4. Indoor data included videos from local herders' breeding areas in Hailar District, Hulunbuir City, and open-source datasets. Through video surveillance in the breeding areas, data were continuously recorded for 3 d from 5 different angles, with 5 min of footage captured every hour, covering both daytime and nighttime. The video format was MP4. These 52 video clips were converted into frame images and saved in JPG format. Images without cows in the scene were manually deleted, and the SSIM (Structural Similarity Index Measure) algorithm[5] was employed to remove redundant images. Data with SSIM similarity parameters exceeding 0.85 were eliminated. From these images and open-source datasets, a total of 3 000 images were selected to construct a daily behavior dataset for cow herds, featuring eight types of daily behaviors: standing, walking, drinking, lying, eating, mounting, fighting, and limping. The dairy cow behavior dataset was divided into training, test, and validation sets at an 8 : 1 : 1 ratio, resulting in a training set comprising 2 400 cow behavior images, a test set comprising 300 images, and validation set comprising 300 images. Fig. 1 illustrates the different behaviors of dairy cows.

Method and Algorithm Design

Improved YOLOv11n network model

To address object detection tasks involving small targets and complex backgrounds, a multi-dilation channel optimization module was applied to the neck network, and a dimension-aware selective fusion module was integrated into the backbone network. These enhancements strengthen the network's feature representation capability, enabling it to extract features at different scales and levels of abstraction while effectively integrating shallow and deep features. It demonstrates strong adaptability for multi-scale object detection, with improved accuracy in localizing and classifying small targets. The model also offers advantages such as a simple and efficient architecture, reduced parameters, and lower computational complexity, which collectively enhance detection speed and optimize feature fusion performance, thereby enabling more effective extraction of complementary feature information.

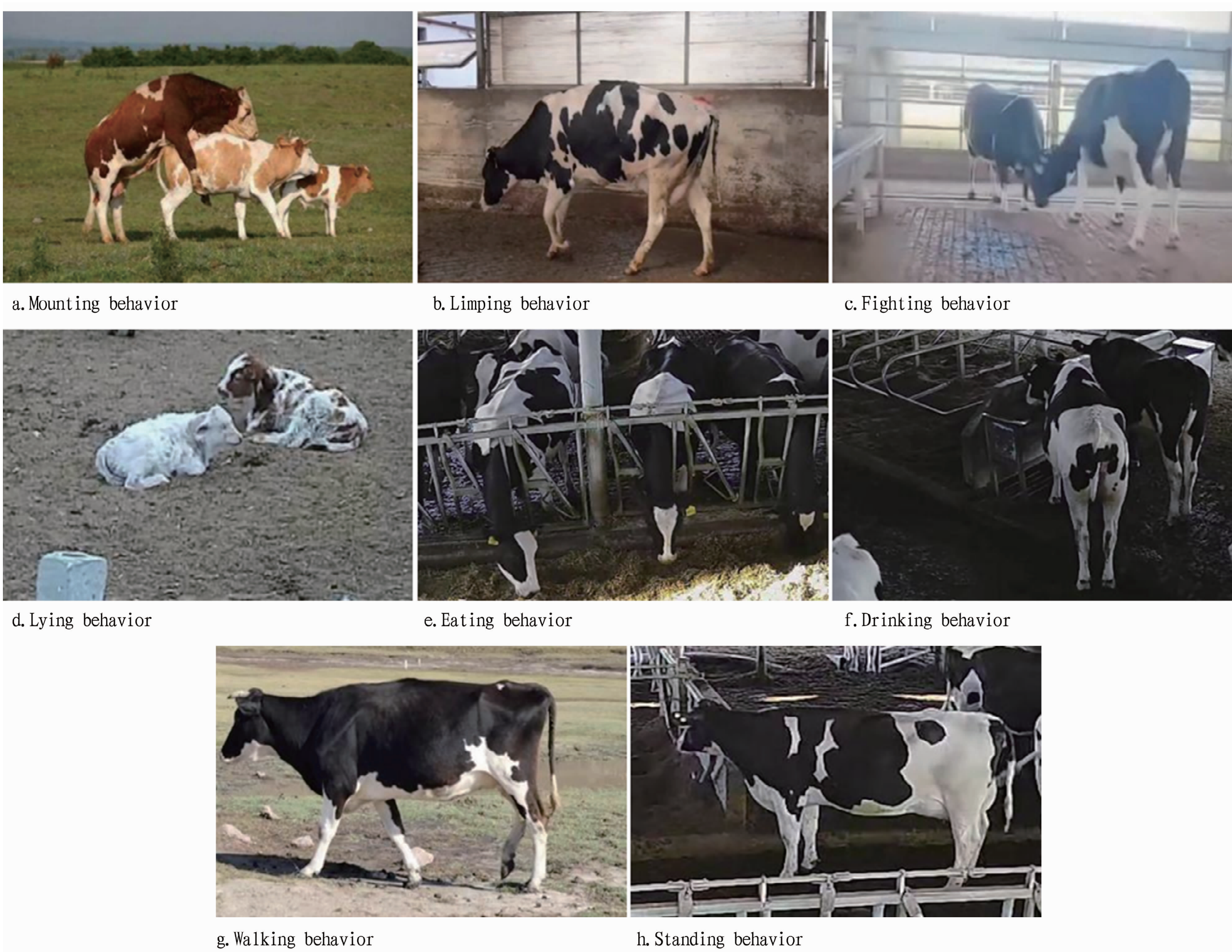

Fig. 1 Examples of dairy cow behavior

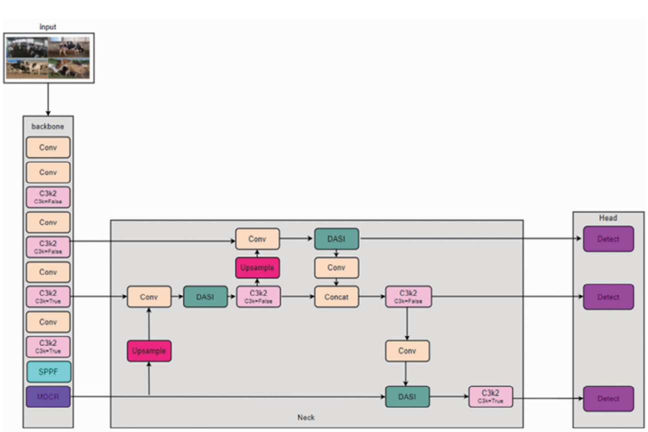

Fig. 2 Improved network structure of YOLOv11

Based on the YOLOv11 network architecture, the model was improved, and the enhanced structure, shown in Fig. 2, can be divided into three parts: Backbone, Neck, and Head. The Backbone adopted the C3k2 block to replace the C2f block and incorporated the DASI module. The Neck performed feature fusion and enhancement and incorporated the MDCR module. The Head included the Silu activation function, adaptive anchor boxes, and multi-scale training.

Dimension-aware selective integration (DASI) module

The DASI (dimension-aware selective integration) module is a key component of the HCF-Net network[6], designed to address the fusion of high-dimensional and low-dimensional features in infrared small target detection. It selects multi-dimensional information from various perspectives based on inter-channel correlations and feature intensity within the network, and integrates them to enhance the visibility of small targets (Fig. 3).

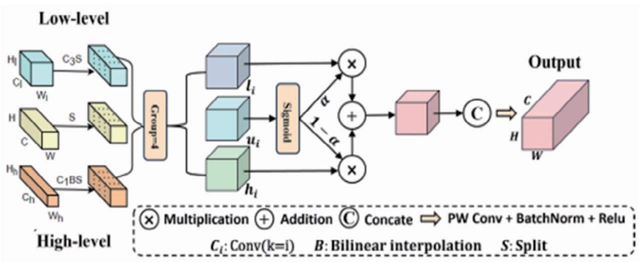

Fig. 3 DASI architecture diagram

The DASI module adaptively selects either high-dimensional or low-dimensional features, effectively preserving detailed information such as the contours and textures of small targets, thereby enhancing their recognizability. Additionally, it adaptively fuses high-dimensional and low-dimensional features based on the size and intrinsic characteristics of small targets. This adaptive selection mechanism enables the module to be applied to various tasks across different environments.

Multi-dilation channel refinement (MDCR) module

The MDCR module employs multi-layer depthwise separable convolutions to capture spatial feature information with varying receptive field sizes, enabling the model to better distinguish infrared small targets[7]. Meanwhile, the input features are split along the channel dimension into several independent heads, each corresponding to a specific feature type. These heads then generate distinct feature representations through independent depthwise separable dilated convolutions.

The input feature $F_a \in R^{H\times W\times C}$ is split into four distinct heads, generating $(a_i)_{i-1}^{4} \in R^{H\times W\times C/4}$. Depthwise separable dilated convolution is applied to each head, resulting in $(a'_i)_{i-1}^{4} \in R^{H\times W\times C/4}$.

$$a'_i = \mathrm{DDWConv}\ (a_i)$$
$$h_j = \mathrm{Winner}([a_1^j,\ a_2^j,\ a_3^j,\ a_4^j])$$
$$F_o = 6(\beta(\mathrm{Wouter}\ ([h_1,\ h_2,\ \cdots,\ h_j])))$$

Here, Winner and Wouter represent the weight matrices used in the pointwise convolution. a'_i is split into individual channels, yielding $(a_i^j)_{c/4}^{j=1} \in R^{H\times W\times 4}$ for each head. These channels are then interleaved to form $(h_j)_{c/4}^{j=1} \in R^{H\times W\times 1}$. Finally, pointwise convolution is applied to facilitate information fusion both within and across groups, producing the output $F_o \in R^{H\times W\times C}$ (Fig. 4).

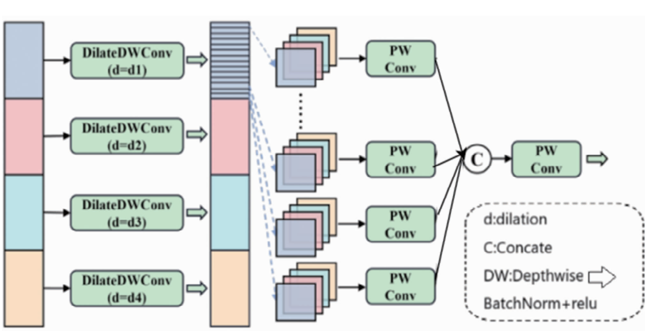

Fig. 4 MDCR architecture diagram

Experiment and Results Analysis

Experimental environment

The experimental platform was Windows 11, with an NVIDIA GeForce RTX 4060 Laptop GPU, a 12th Gen Intel(R) Core(TM) i7-12650H processor, and 32 G of RAM. The virtual environment was set up using Anaconda 3. The IDE environment employed PyCharm Community. The open-source framework utilized PyTorch. The programming language was Python 3.10. The training was performed with an initial learning rate of 0.01 and a batch size of 4 for a total of 200 training epochs. The input image size was configured to 640 × 640 pixels, and a stochastic gradient descent (SGD) optimizer was selected.

Data ablation experiment

From the experimental results, it could be observed that YOLOv11n + DASI + MDCR achieved an overall improvement of 3.5 percentage points in mean average precision (mAP) compared with YOLOv11n. Among individual behaviors, the performance for lying was slightly lower than that of YOLOv11n and other improved models, while all other behaviors achieved the highest performance. Specifically, the performance was improved by 2.8 percentage points for standing behavior, and by 2.0 percentage points for eating behavior, and the improvements for drinking and walking were most significant, reaching 8.7 and 6.3 percentage points, respectively. In terms of drinking, contour information predominates in trough targets, while texture information is relatively scarce. Additionally, due to camera angle issues, the trough is often obstructed by the cow's head, leading to lower recognition accuracy of the trough. As for walking behavior, its similarity to limping is high. The improved model provides greater distinction between these two behaviors, resulting in more noticeable recognition effect. This demonstrates that the addition of the DASI and MDCR modules enhances the detailed recognition of cows drinking, particularly in capturing the subtle features of a cow's head at the water trough (Table 1).

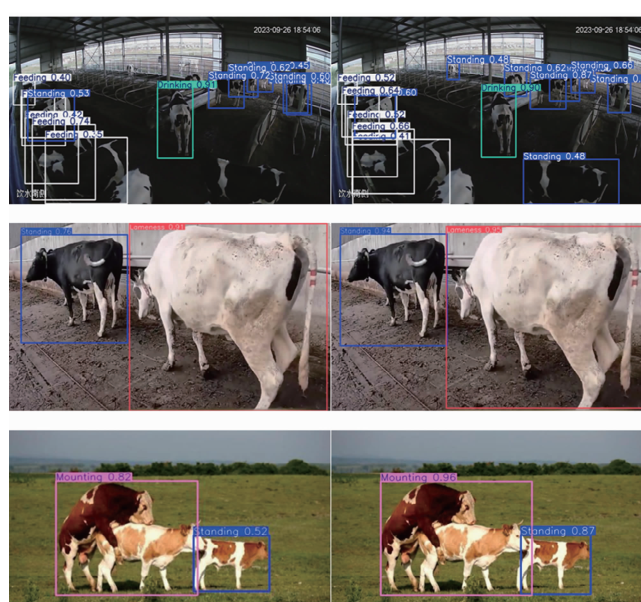

Fig. 5 Comparison of detection effect

Table 1 Results of the ablation comparison experiment

Model	Mean average precision (mAP)	Average precision (AP)							
		Standing	Lying	Eating	Drinking	Fighting	Mounting	Limping	Walking
YOLOv11n	0.895	0.773	0.876	0.908	0.780	0.961	0.990	0.967	0.907
YOLOv11n-RCM	0.905	0.781	0.909	0.912	0.740	0.977	0.994	0.978	0.946
YOLOv11n-CBM	0.909	0.795	0.927	0.918	0.751	0.980	0.995	0.970	0.932
YOLOv11n-DASI	0.916	0.751	0.913	0.928	0.806	0.985	0.994	0.981	0.967
YOLOv11n-DASI-MDCR	0.930	0.801	0.907	0.928	0.867	0.988	0.995	0.987	0.970

Analysis of model training results

To evaluate the performance optimization and improvement effects of the model, the basic YOLOv11n model and the improved YOLOv11n-DASI-MDCR model were selected for dual comparative experiments under different scenarios in this study. Both indoor and outdoor conditions were included in the comparison, while high-density gatherings of cow herds under real pasture working conditions and varying lighting intensity conditions were used as the primary simulation environments. From the quantitative comparison results in Fig. 5, it is evident that in multi-target scenarios characterized by overlapping objects and uneven lighting, the improved model better preserves feature information while simultaneously increasing the confidence levels for recognizing key cow behaviors. Notably, it demonstrates enhanced performance in localization accuracy under low-light conditions (such as dawn and dusk) and in dense cow areas.

Conclusions

To address the challenges of multi-target cow behavior recognition under complex environmental conditions, an intelligent detection method based on an improved YOLOv11n model was proposed. By integrating a DASI module into the backbone network and an MDCR module into the neck network, this method significantly enhances the model's adaptability to small targets and complex backgrounds. Notably, it achieves substantial breakthroughs in recognizing key cow behaviors such as drinking and walking. The DASI module effectively addresses feature information loss and background clutter in small target detection through adaptive feature selection and fusion, while the MDCR module enhances small target localization capability by capturing spatial features of different receptive fields through multiple depthwise separable convolutional layers. Experimental results indicate that the improved YOLOv11n-DASI-MDCR model increased the accuracy rate from the original 85.4% to 90.7%, and the mean average precision (mAP) rose from the original 89.5% to 93%. The recognition accuracy for drinking and walking behaviors was improved by 8.7% and 6.3%, respectively, effectively addressing issues related to occlusion caused by camera angles and interference from similar behaviors. Furthermore, the model demonstrated strong robustness in various scenarios, including indoor and outdoor environments, varying lighting conditions, and dense cattle gatherings.

References

[1] WANG Z, ZONG ZY, WANG HC, *et al.* Research status and development of digital detection methods for cow estrus[J]. China Feed, 2021(21): 134–138. (in Chinese).
[2] HAO Y, JIANG XJ, SONG YX, *et al.* Investigation report on limb and hoof disease of dairy cows in 87 large-scale pastures[J]. Chinese Journal of Veterinary Science, 2024(12): 1–9. (in Chinese).
[3] RUMELHART DE, HINTON GE, WILLIAMS RJ. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088): 533–536.
[4] DU YR. Research on daily behavior recognition of cattle based on computer vision[D]. Baotou: INNER Mongolia University of Science & Technology, 2022. (in Chinese).
[5] AFSHIN TATAR, MANOUCHEHR HAGHIGHI, ABBAS ZEINIJAHROMI. Experiments on image data augmentation techniques for geological rock type classification with convolutional neural networks[J]. Journal of Rock Mechanics and Geotechnical Engineering, 2025, 17(1): 106–125. (in Chinese).
[6] QIN YM, YIN LJ, GAO XN, *et al.* Infrared image hot spot fault detection for photovoltaic modules using HCF-YOLO[J]. Acta Energiae Solaris Sinica, 2025(2): 1–9. (in Chinese).
[7] XU SB, ZHENG SC, XU WH, *et al.* HCF-Net: Hierarchical context fusion network for infrared small object detection[C]//2024 IEEE International Conference on Multimedia and Expo. Canada: Niagara Falls Marriott on the Falls, 2024: 1–6.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU

(Continued from page 35)

[4] SUN Y, WANG JY, LYU J. Application of antibody detection technology in animal disease prevention and control[J]. China Animal Health, 2023, 25(10): 11–12.
[5] GAO DL. Key prevention and treatment measures for spring swine diseases[J]. Cereals, Oils and Feed Science and Technology, 2024(11): 143–145.
[6] LI P. Construction and research on the evaluation indicator system for major animal disease status[D]. Inner Mongolia Agricultural University, 2014.
[7] FANG ZR. Integrated prevention and control strategies for respiratory diseases in large-scale pig farms[J]. China Animal Health, 2025, 27(8): 4–6.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU