Novel method for the recognition of Jinnan cattle action using bottleneck attention enhanced two-stream neural network

Wangli Hao[1], Meng Han[1], Kai Zhang[1], Li Zhang[1], Wangbao Hao[2,3], Fuzhong Li[1], Zhenyu Liu[4*]

(1. *School of Software, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China*;
2. *Yuncheng National Jinnan Cattle Genetic Resources and Gene Protection Center, Yongji 044500, Shanxi, China*;
3. *Yongji Puzhou Old City Beimenwai Cattle Farm, Yongji 044500, Shanxi, China*;
4. *College of Engineering, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China*)

Abstract: Effective and accurate action recognition is essential to the intelligent breeding of the Jinnan cattle. However, there are still several challenges in the current Jinnan cattle action recognition. Traditional methods are based on manual characteristics and low recognition accuracy. This study is aimed at the efficient and accurate development of Jinnan cattle action recognition methods to overcome existing problems and support intelligent breeding. The acquired data from the previous methods contain a lot of noise, which will cause individual cattle to have excessive behaviors due to unsuitability. Concerning the high labor costs, low efficiency, and low model accuracy of the above approaches, this study developed a bottleneck attention-enhanced two-stream (BATS) Jinnan cattle action recognition method. It primarily comprises a Spatial Stream Subnetwork, a Temporal Stream Subnetwork, and a Bottleneck Attention Module. It can capture the spatial-channel dependencies in RGB and optical flow two branches respectively, so as to extract richer and more robust features. Finally, the decision of the two branches can be fused to gain improved cattle action recognition performance. Compared with the traditional methods, the model proposed in this study has achieved state-of-the-art recognition performance, and the accuracy of motion recognition was 96.53%, which was 4.60% higher than other models. This method significantly improves the efficiency and accuracy of behavior recognition and provides an important research foundation and direction for the development of higher-level behavior analysis models in the future development of smart animal husbandry.

Keywords: Jinnan cattle, action recognition, bottleneck attention, two-stream neural network
DOI: 10.25165/j.ijabe.20241703.8202

Citation: Hao W L, Han M, Zhang K, Zhang L, Hao W B, Li F Z, et al. Novel method for the recognition of Jinnan cattle action using bottleneck attention enhanced two-stream neural network. Int J Agric & Biol Eng, 2024; 17(3): 203–210.

1 Introduction

Effective cattle action recognition plays a vital role in understanding the life habits of cattle. Further, it is also important in improving the quality of beef and the breeding of cattle. Cattle action recognition has been widely utilized in various fields of intelligent breeding, which is conducive to the development of intelligent breeding and precision breeding.

Recently, with the development of artificial intelligence technologies and the improvement of computer hardware performance, artificial intelligence technologies have been widely employed in numerous fields such as education, medicine, business, and agriculture respectively, which have also been utilized in animal husbandry to improve feeding management. Conventionally, this management has been carried out through human observation. However, this approach needs lots of time and effort, but with low accuracy.

Sensor technologies, capable of automatically monitoring action patterns, have been utilized for cattle[1], but several of these sensors are harmful and may cause cattle stress. Additionally, there are additional costs associated with the utilization of sensors. In recent, livestock action recognition, which leverages image classification with visual object detection, has become a topic of increasing interest and has been an increasingly popular topic in the fields of computer vision and smart agriculture[2,3], respectively.

Computer vision techniques that permit non-attached observations of cattle have recently gained increasing attention in cattle action recognition. This research generally depends on monitoring the estrus state of cattle via identifying actions such as “riding”[4,5]. The research on action recognition through deep learning methods[6-8] has achieved great success.

Furthermore, according to the observations, some salient regions can clearly represent the actions of the cattle. For example, head features can allow us to recognize if they are grazing, and the body parts can clearly show whether they are standing or lying. Focusing on these areas may improve the recognition performance. Therefore, an attention mechanism was adopted for the backbone recognition network to get more powerful and discriminative recognition features.

Ahn et al.[9] proposed a cattle behavior recognition algorithm,

Received date: 2023-02-21 **Accepted date:** 2024-05-17

Biographies: Wangli Hao, PhD, Associate Professor, research interest: computer vision, pattern recognition, and machine learning, Email: haowangli@sxau.edu.cn; **Meng Han**, PhD candidate, research interest: distributed and parallel computing, distributed machine learning, Email: hanm@hdu.edu.cn; **Kai Zhang**, Master candidate, research interest: smart agriculture, Email: hualimengyu@163.com; **Li Zhang**, Master candidate, research interest: deep learning applications in smart agriculture, Email: z20213621@stu.sxau.edu.cn; **Wangbao Hao**, Breeder, research interest: breeding and precision feeding of Jinnan cattle, Email: 59742516@qq.com; **Fuzhong Li**, PhD, Professor, Doctoral supervisor, the Dean of the Software College of Shanxi Agricultural University, Email: lifuzhong@sxau.edu.cn.

***Corresponding author: Zhenyu Liu**, PhD, Professor, Doctoral supervisor, Postdoctoral Researcher at China Agricultural University, the deputy Dean of the Agricultural Engineering College, Shanxi Agricultural University. research interest: electromagnetic properties of agri-cultural materials and livestock informatization. Tel: +86-354-6287098, Email: lzysyb@126.com.

which utilized a support vector machine (SVM) classifier with feature information of motion history images to detect the optimal insemination time. Wu et al.[10] proposed an effective corbel detection method based on YOLOv3 and relative step size characteristic vector. Specifically, based on the relative step size characteristic vector, the YOLOv3 algorithm was utilized to detect the position of the corbel, and then the LSTM model was employed to identify the normal walking and the lame behavior of the cattle, with an accuracy of 98.57% obtained.

To overcome the problems brought by manual lameness detection, such as difficulty in detecting sudden, severe, or early lameness, Jiang et al.[11] proposed to utilize a neural network with single-stream long-term optic flow convolution, to learn video representation flow convolutions. Shen et al.[12] first applied the YOLO model to detect cattle, and then an improved AlexNet model was employed to classify the corresponding detected individual cattle. Finally, they obtained 96.65% accuracy in individual cattle classification. Tassinari et al.[13] proposed a deep learning-based system for individual cattle classification and location analysis. Zhang et al.[14] proposed a lightweight YOLO detection model, using MobileNetV3 to replace the backbone network in the YOLOv3 network, and obtained 96.8% cattle key position detection accuracy. Hu et al.[15] employed the YOLO algorithm to extract cattle objects, and then a segmentation algorithm was utilized to extract the head, torso, and leg parts of the corresponding cattle object. Subsequently, deep feature fusion was performed on these extracted parts. Finally, the SVM classifier was employed to do the classification, and an accuracy of 98.36% was obtained. Jiang et al.[16] proposed a filter-based YOLOv3 algorithm and achieved 99.18% accuracy in detecting key parts of cattle. Based on an RGB camera and convolutional neural network, Bezen et al.[17] built a computer vision system for measuring cattle feeding, and an accuracy of 93.65% was obtained. Achour et al.[18] has built a CNN-based image analysis system for the classification of individual cattle, their foraging behavior, and the food respectively. Specifically, their model obtained an accuracy of 97.00% for individual cattle classification and an accuracy of 92.00% for cattle foraging behavior separately. Wu et al.[19] proposed a Fusion of Convolutional Neural Network and Long Short-Term Memory Network (CNN-LSTM) model for cattle action recognition. Specifically, the action categories of cattle in their experiments included drinking, ruminating, walking, standing, and lying down, respectively, and the average classification accuracy of their model reached 97.6%. Xue et al.[5] developed a model to perform cattle feeding, lying, and standing action recognition tasks. They first employed the YOLOv5 algorithm to detect cattle, then constructed the spatial relationship vector of the feature parts based on detected figures, and finally neural networks were employed to conduct the classification task.

He et al.[6] proposed a novel approach for cattle action recognition based on video analysis. Specifically, the authors first initialized the image background, performed the preprocessing of the median filter difference on each input frame, and then conducted the gamma transformation on the preprocessed image. The target images were segmented based on the gray-scale feature segmentation method, the entire target was cyclically searched through the largest connected area, and the backgrounds were updated in real-time. Consequently, a fast clustering algorithm was implemented based on the extracted centroid variance and contour features to realize the action recognition with six categories for cattle, including lying, standing, walking, running, and jumping respectively. Xie et al.[7] developed a novel cattle crawling behavior recognition approach based on machine vision. Firstly, the input videos were split into several segments, and then three features of width, height, and aspect ratio of the smallest bounding rectangle were extracted. Consequently, the relationships between these three features and time were established separately. Specifically, time series motion curves of width, height, and aspect ratio were drawn to simulate the behavior of cattle. Finally, backpropagation (BP) neural network and k-nearest neighbor (KNN) were utilized to realize the effective action recognition of cattle crawling behavior.

As above stated, although the mentioned cattle action recognition methods have achieved promising performance, several problems still exist. For example, there are many parameters and the training is difficult. Moreover, the features they extracted rely only on the RGB image in the spatial domain whilst ignoring the temporal domain. To tackle these problems, this study developed a novel cattle action recognition approach based on bottleneck attention-enhanced two-stream neural network for Jinnan cattle action recognition. Specifically, two branch networks were constructed, one was employed to extract RGB spatial features, and the other one was utilized to extract the optical flow features in the temporal domain. In this way, the mole of the study of this paper can capture both spatial and temporal information of input videos and avoid the accuracy decreased problem caused by only spatial domain features extracted in a single network. Additionally, the model of this study integrates bottleneck attention modules in the two branch networks respectively. This bottleneck attention module can build spatial-channel dependencies, which is conducive to extracting more robust and rich features to obtain superior cattle action recognition performance.

Furthermore, in order to validate the effectiveness of the model and provide materials and a foundation for intelligent breeding researchers, a new cattle action recognition dataset was established in this study. It contains six action categories, including drinking, eating, looking back, lying, walking, and standing, respectively. Each action category contains about 300 samples, and the video duration varies from 5 to 10 s.

Above all, our contributions can be summarized as follows:

1) First, a novel Jinnan cattle action dataset was established, which contained 6 categories. Each category comprises about 300 videos and each video ranges from 5 to 10 s. The original videos were collected from eight cameras on a farm for more than a month. There are more than 200 cattle on the farm and the age ranges from newborn calves to old cattle. All the above factors lead to the novel established dataset collected on this farm having better diversity in terms of illumination, age, angle, etc.

2) A novel cattle action recognition method was proposed based on bottleneck attention enhanced two-stream neural network. This model can establish the spatial-channel dependencies in RGB and Flow two paths separately and will extract both appearance and temporal information with richer and more robust features. Simultaneously, the decisions of the RGB and Flow two branches can be merged to gain an improved cattle action recognition performance. Specifically, the model proposed in this study achieved the best performance for the Jinnan cattle action recognition task, with about 5% improvement of the existing model.

3) Numerous ablation experiments have been designed and performed to verify the performance of our model. Specifically, these studies include the evaluation of the superiority of the proposed model, the evaluation of the effectiveness of the two-stream network, and the evaluation of the effectiveness of the attention module respectively.

2 Materials and methods

2.1 Datasets

The Jinnan cattle action recognition dataset utilized in this study was collected from one farm in Yongji City, Shanxi Province, containing a total of 227 Jinnan cattle specimens. The farm is 90.0 m long and 20.0 m wide, with 5 enclosures in total. The cameras were installed at the fixed position of each enclosure. Specifically, each enclosure contained four cameras, distributed in different places, including 2.5 m above the enclosure sports ground and water tank, and 2.5 m above the 2 ends of the food tank, respectively. Thus, the distribution of the camera can facilitate the observation of the feeding, drinking, and other behaviors of cattle. The videos were collected in February 2021, lasting for 1 month. Under natural light conditions, the video of cattle activities on the farm from 7 a.m. to 6 p.m. was collected. Seven behaviors of calves, including standing, walking, lying, eating, drinking, and looking back, respectively, were collected. The quantity of Jinnan Cattle in different movements is as follows: there are 295 heads of standing, 303 heads of walking, 279 heads of lying, 342 heads of eating, 302 heads of drinking, and 300 heads of looking back. These different action states of Jinnan Cattle have reached 1821 heads, which provides us with important references on the behavior habits of Jinnan Cattle.

After removing invalid clips with poor quality, such as motion blur and bad weather, a total of 2 TB videos were collected. 300 clips of each behavior were filtered out by manually intercepting videos, each of which lasted about 5-15 s. The collected videos were stored in MP4 format, with a video frame rate of 25 fps. An example of the collected data is shown in Figure 1.

In order to enhance the generalization ability of the model, this study expanded the collected data based on the data augmentation. Data augmentation refers to increasing the number and the diversity of the input samples by rotating and cutting existing data. Furthermore, this operation can alleviate the overfitting of the model. Specifically, this study employed rotation, flip, clipping, deform, scaling, shift, color jitter, and other approaches to perform the data augmentation.

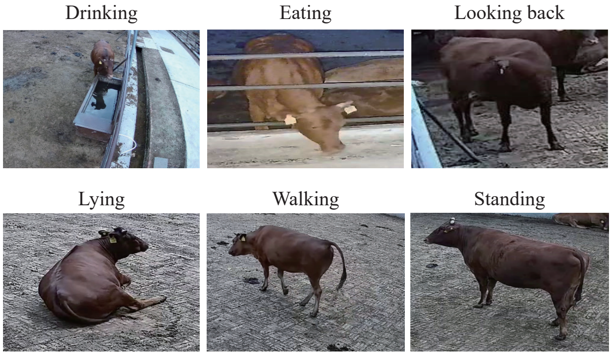


Figure 1 Examples of various actions of Jinnan cattle, including drinking, eating, looking back, lying, walking, and standing respectively

2.2 Problem definition

The proposed BATS model for Jinnan cattle action recognition was formulated as shown in Figure 2. Assume there are N videos $\chi = \{x_i\}_{i=1}^{N}$ in the dataset belonging to class C, and the corresponding video action label set is described as $\gamma = \{y_i\}_{i=1}^{N}$ with $y_i \in \{1, 2, 3, \ldots, C\}$.

The probability $p_s^c(x_i^s)$ of the random selected RGB image x_i^s from video x_i is predicted to be class c in the RGB spatial stream network can be gained by

$$p_s^c(x_i^s) = \frac{\exp\left(S_\tau^c(x_i^s)\right)}{\left(\sum_{c=1}^{C} \exp\left(S_\tau^c(x_i^s)\right)\right)} \tag{1}$$

where, x_i represents the ith video, c stands for one class in class C, x_i^s indicates the randomly selected RGB image from the video x_i, and $S_\tau^c(x_i^s)$ denotes the logit value obtained from the softmax layer of the spatial stream network for the RGB image x_i^s of video x_i.

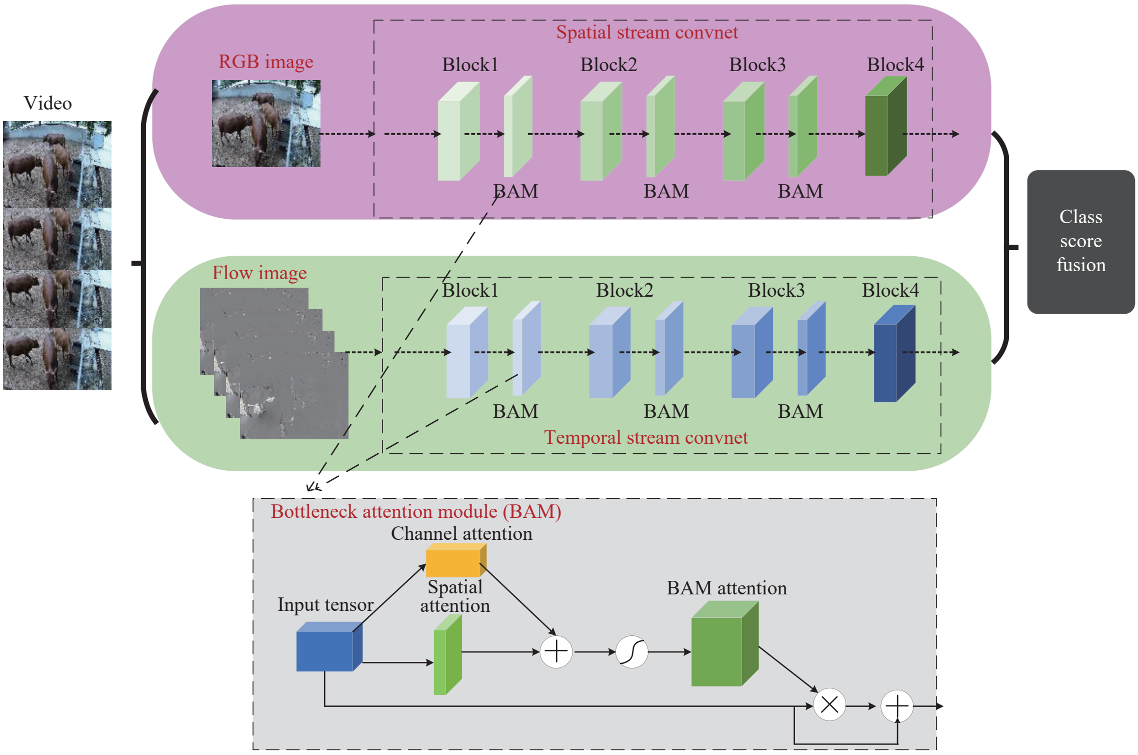


Figure 2 Basic pipeline of the proposed bottleneck attention and two-stream neural network (BATS) model for Jinnan cattle action recognition

The probability $p_t^c(x_i^t)$ of the optical flow image x_i^t corresponding to RGB image x_i^t from video x_i is predicted to be class c in the flow temporal stream network can be obtained by

$$p_t^c(x_i^t)=\frac{\exp\left(S_t^c(x_i^t)\right)}{\left(\sum_{c=1}^{C}\exp\left(S_t^c(x_i^t)\right)\right)} \tag{2}$$

where, x_i^t indicates the flow image corresponding to x_i^s at the same time from video x_i, $S_t^c(x_i^t)$ represents the logit value obtained from the softmax layer of temporal stream network.

The loss function for the whole network is defined as:

$$\varsigma=(1-\alpha)\varsigma_s+\alpha\varsigma_t \tag{3}$$

where, ς_s denotes the loss term of the spatial stream neural network, ς_t indicates the loss term of the temporal stream neural network respectively, and α is the hyperparameter employed to balance the two loss terms. Specifically, the ς_s and ς_t represent the traditional cross-entropy losses and can be achieved by the following equations separately.

$$\varsigma_s=-\sum_{i=1}^{N}\sum_{c=1}^{C}y_i^c\log\left(p_s^c(x_i^s)\right) \tag{4}$$

$$\varsigma_t=-\sum_{i=1}^{N}\sum_{c=1}^{C}y_i^c\log\left(p_t^c(x_i^t)\right) \tag{5}$$

2.3 Bottleneck attention and two-stream neural network architecture for Jinnan cattle action recognition

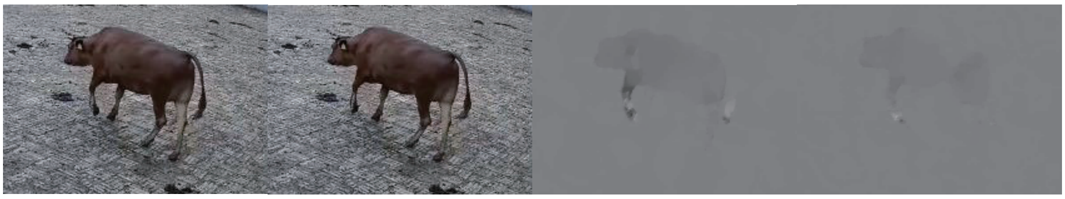
In this section, the proposed model Bottleneck attention enhanced two-stream (BATS) is depicted in detail (the pipeline is presented in Figure 2). BATS mainly contains two subnetworks, including the spatial stream neural network and the temporal stream neural network respectively. The purpose of it is to explore the spatial-channel dependencies among appearance and motion streams underlying the source video.

The proposed model was built on the top of the two-stream networks[20], which contain RGB and Flow subnetworks respectively. The reasons were that video can naturally be divided into the spatial and temporal parts. Concerning the spatial part, it focuses on the individual frame appearance that implies information about objects and scenes described in the video. For the temporal part, it concentrates on the motion across the frames, which carries the movement between the objects and the observer (the camera). Each subnetwork is implemented based on a deep convolutional network, and the softmax scores of them are finally fused. The only difference between the two subnetworks is their input dimensions, which depend on the channel of the input image. Concretely, the input dimension is 3 for the RGB stream and 2 for the Flow stream, separately. It should be noted that the basic neural network for two streams in our experiment is ResNet.

Based on the two-stream network, the proposed model BATS leverages the bottleneck attention module into the framework (as shown in Figure 2). This bottleneck attention module can capture the channel-spatial attention, so as to extract more abundant features, thus improving the accuracy of the Jinnan cattle action recognition. To obtain the hierarchical channel-spatial attention, this BAM module is integrated at each bottleneck of basic models where it performs the down-sampling of feature maps.

1) Spatial stream subnetwork: Spatial stream convolutional network (ConvNet) focuses on individual video frames, effectively executing Jinnan cattle action recognition from still images. The static appearance itself is a useful clue because some actions are strongly associated with particular objects. Since the spatial stream network is an image classification framework, in essence, the network was pre-trained on the large-scale ImageNet[21] dataset to obtain excellent action recognition performance.

2) Temporal stream subnetwork: The temporal stream convolutional network operates on the motion between video frames, effectively performing Jinnan cattle action recognition from several consecutive frames. Specifically, the input to the temporal stream subnetwork in this study was formed by stacking two directional optical flow displacement fields between several consecutive frames. Such input clearly describes the motion between video frames, which makes action recognition easier. The sample flow images for the consecutive video frames utilized in this study are presented in Figure 3. Figures 3a and 3b show the two adjacent images. Figure 3c is the optical flow image of two adjacent frames in the horizontal direction; Figure 3d is the optical flow image of two adjacent frames in the vertical direction.



a. b. c. d.

Note: a and b are two adjacent images; c is the optical flow image of two adjacent frames in horizontal direction; d is the optical flow image of two adjacent frames in vertical direction.

Figure 3 Sample flow images for the consecutive video frames

3) Bottleneck attention module (BAM): The BAM[22] module is utilized for capturing the spatial and channel dependencies underlying the video image. Its detailed architecture is shown in Figure 2. Based on an intermediate input tensor $F\in R^{C\times H\times W}$, a 3D attention map $M(F)\in R^{C\times H\times W}$ contains spatial and channel information can be gained. Moreover, the whole attention process can be represented as:

$$F'=F+F\otimes M(F) \tag{6}$$

where, $\otimes$ denotes the element-wise multiplication, F' indicates the final processed feature map. Additionally, $M(F)$ can be acquired by the following equation:

$$M(F)=\mathrm{Sigmoid}(M_c(F)+M_s(F)) \tag{7}$$

where, $M_c(F)\in R^{C}$ represents the channel attention feature map and $M_s(F)\in R^{C\times W}$ indicates the spatial attention feature map separately. It should be noted that $M_s(F)$ and $M_c(F)$ are all expanded to the same size as that of F before addition.

Among them, $M_c(F)$ and $M_s(F)$ can be achieved through the following definitions:

$$M_c(F) = \mathrm{BN}\left(\mathrm{MLP}\left(\mathrm{AvgPool}(F)\right)\right) \tag{8}$$

$$M_s(F) = \mathrm{BN}\left(f_3^{1\times1}\left(f_2^{3\times3}\left(f_1^{3\times3}\left(f_0^{1\times1}(F)\right)\right)\right)\right) \tag{9}$$

where, BN indicates a batch normalization operation, f denotes a convolution operation, the subscripts of f index the index of the convolutional layer, and the superscripts of f represent the kernel size of the convolution layer. The two 1×1 convolutions are employed for channel reduction and the two 3×3 dilated convolutions are applied to aggregate the contextual information through a larger receptive field.

The more detailed computation and concatenation processes of $M_c(F)$ and $M_s(F)$ can be found in the study of Reference [22].

2.4 Implementation details

For a fair comparison, all experiments were implemented and run on the PyTorch framework. The stochastic gradient descent algorithm (SGD) was utilized to train models and the minibatch size was 16. All comparison models were initialized by the pre-training model based on ImageNet[52]. The learning rate started at 0.001 and then decreased to 1/10 per 200 epochs, and the value of momentum was 0.9. The total training was 500 epochs.

The equipment employed in this study is set as follows: the operating system was Ubuntu 18.04, the CPU was InterTMi7-7800X, the GPU was NIVIDA TITAN Xp, the memory was 16× 6 GB DDR4, the mechanical hard disk was 4 TB, and the solid state disk was 512 GB.

3 Experimental results and analysis

In this section, the experimental results and the analysis will be reported in detail. The whole experiment was comprised of the following several design parts, including the evaluation of the superiority of the proposed model, the evaluation of the effectiveness of the two-stream network, and the evaluation of the effectiveness of the attention module correspondingly.

3.1 Evaluation of the superiority of the proposed model

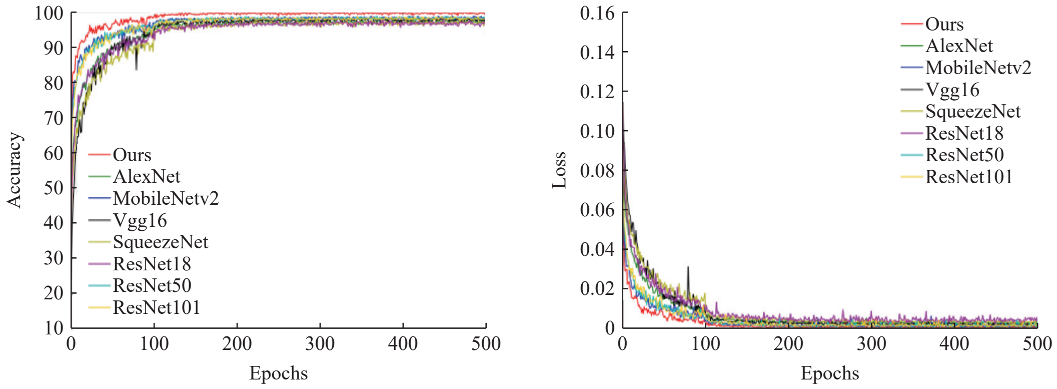
In order to validate the superiority of the proposed BATS, several models were utilized for comparison, including AlexNet[21], MobileNetV2[23], Vgg16[24], SqueezeNet[25], ResNet18[26], ResNet50[26] and ResNet101[26], respectively. The results are listed in Table 1.

Table 1 Compared accuracies of different models

Model	Accuracy/%
AlexNet	92.14
MobileNetv2	93.29
Vgg16	93.57
SqueezeNet	93.00
ResNet18	92.28
ResNet50	93.71
ResNet101	95.10
The proposed method in this study	**96.53**

Table 1 lists that the proposed model obtains a better Jinnan action recognition performance than other common models. Specifically, the proposed model achieved 96.53% accuracy, which was 4.76%, 3.47%, 3.16%, 3.80%, 4.60%, 3.00%, and 0.15% better than those of AlexNet, MobileNetv2, Vgg16, SqueezeNet, ResNet18, ResNet50, ResNet101, correspondingly. These results validate the superiority of the proposed model.

Moreover, in order to allow readers to have a more intuitive perception of the superiority of BATS in Jinnan cattle action recognition task, the accuracies and the losses of different comparison models under different epochs are reported in Figure 4a is the accuracy for different models under different epochs. Figure 4b is the loss for different models under different epochs. Figure 4 shows that the accuracy and the loss of BATS exceed those of other models, which validates the effectiveness of the proposed BATS.



a. Accuracy for different models under different epochs

b. Loss for different models under different epochs

Note: Ours means the proposed method in this study.

Figure 4 Comparison results of different models in terms of accuracy and loss

The advantages of the novel model are attributed to its ability to capture the spatial-channel dependence on the appearance and optical flow and extract richer and more robust features, thus improving the performance of Jinnan cattle action recognition.

3.2 Evaluation of the two-stream network in the Jinnan cattle action recognition

Concerning the validation of the effectiveness of the two-stream network setting in the Jinnan cattle action recognition framework, the single RGB stream, single Flow stream, and the fusion of two streams (simplified as Two-steam in Table 2) of several models are utilized for comparison. The compared models include AlexNet[21], MobileNetv2[23], Vgg16[24], SqueezeNet[25], ResNet18[26], ResNet50[26] and ResNet101[26] separately. The results are listed in Table 2.

Table 2 lists that the two-stream network setting achieved a significantly consistently superior performance than those of the single RGB and Flow network.

Concretely, the two-stream version of the AlexNet model achieved the 92.14% accuracy, which was 2.87% and 23.33% better than those of its corresponding RGB and flow version; the two-stream version of the MobileNetv2 model achieved the 93.29% accuracy, which was 1.25% and 12.38% better than those of its corresponding RGB and Flow version; the two-stream version of the Vgg16 model achieves the 93.57% accuracy, which was 2.18% and 22.42% better than those of its corresponding RGB and Flow streams; the Two-stream version of the SqueezeNet model achieves the 93.00% accuracy, which was 0.76% and 21.46% better than those of its corresponding RGB and flow version; the two-stream version of the ResNet18 model achieves the 92.28% accuracy, which is 0.62% and 27.65% better than those of its corresponding RGB and Flow streams; the Two-stream version of the ResNet50 model achieves the 93.71% accuracy, which is 0.91% and 25.66% better than those of its corresponding RGB and Flow version; the Two-stream version of the ResNet101 model achieves the 95.10% accuracy, which is 2.24% and 21.47% better than those of its corresponding RGB and flow streams, respectively.

Table 2 Compared accuracies of different models

Model	RGB/%	Flow/%	Two-stream/%
AlexNet	89.57	74.71	92.14
MobileNetv2	92.14	83.00	93.29
Vgg16	91.57	76.43	93.57
SqueezeNet	92.29	76.57	93.00
ResNet18	91.71	72.29	92.28
ResNet50	92.86	74.57	93.71
ResNet101	93.01	78.29	95.10

These comparison results validate the superiority of the two-stream network in the Jinnan cattle action recognition task. The reasons were that the two-stream network can capture both the appearance and the motion information in the video, so the effective spatiotemporal features are extracted, which facilitates the promotion of the performance of Jinnan cattle behavior recognition.

3.3 Evaluation of the effectiveness of the bottleneck attention module in the Jinnan cattle action recognition

In order to verify the effectiveness of the bottleneck attention module in Jinnan cattle action recognition settings, several models were utilized for comparison. They are Res18 and BAMRes18, Res50 and BAMRes50, Res101 and BAMRes101, respectively. Among them, the Res18, Res50, and Res101 refer to the two-stream action recognition model with the basic subnetwork of Renset18, ResNet50, and ResNet101. The BAMRes18, BAMRes50, and BAMRes101 share a similar structure with that of Res18, Res50, and Res101, and the difference between them is that the basic subnetwork of BAMRes18, BAMRes50, and BAMRes101 integrates the Bottleneck attention modules. The comparison results are listed in Table 3.

Table 3 Compared accuracies of different models with or without bottleneck attention module

Model	RGB/%	Flow/%	Two-stream/%
Res18	91.71	72.29	92.28
BAMRes18	**92.18**	**73.86**	**93.73**
Res50	92.86	74.57	93.71
BAMRes50	**93.22**	**77.14**	**95.14**
Res101	93.01	78.29	95.10
BAMRes101	93.42	87.71	**96.53**

Table 3 lists that the bottleneck attention-enhanced two-stream network models achieve better performance. Specifically, BAMRes18 received 93.73% accuracy, which was 1.57% higher than that of Res18; BAMRes50 obtained 95.14% accuracy, which was 1.50% superior to that of Res50; BAMRes101 achieved 96.53% accuracy, which is 1.48% higher than that of Res101. These results verify the superiority of the bottleneck attention module in the two-stream Jinnan cattle recognition framework.

Moreover, the BAM module enhanced single stream network also exhibited superior performance than its corresponding counterpart that was without BAM unit. Concretely, the RGB and Flow stream of BAMRes18 model achieved 0.51% and 2.17% better performance than those of Res18; the RGB and Flow stream of BAMRes50 model achieved 0.39% and 3.45% superior accuracy than those of Res50; the RGB and Flow stream of BAMRes101 model achieved 0.44% and 12.03% superior accuracy than those of Res101.

Furthermore, to offer readers a more intuitive perception of the superiority of the BAM module in Jinnan cattle action recognition task, the accuracies and the losses of different comparison models with or without BAM units under different epochs are illustrated in Figure 5 and Figure 6.

Specifically, the first row of Figure 5 reports the accuracy of the RGB/Flow/Fusion stream on the Res18 and BAMRes18 models under different epochs; the middle row of Figure 5 indicates the Accuracy of the RGB/Flow/Fusion stream on the Res50 and BAMRes50 models under different epochs; the third row of Figure 5 denotes the Accuracy of the RGB/Flow/Fusion stream on the Res101 and BAMRes101 models under different epochs, respectively.

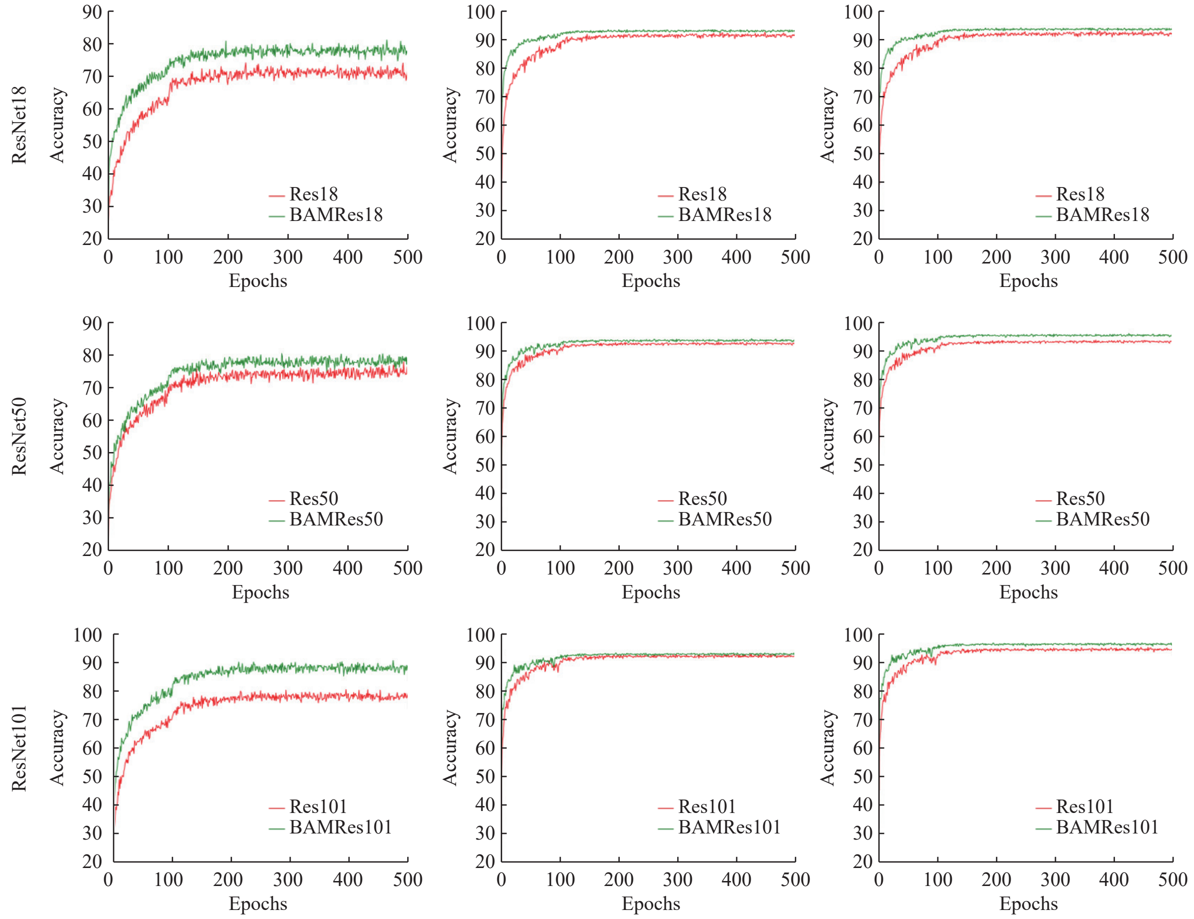
The first column of Figure 6 denotes the Loss of the RGB/Flow stream of the Res18 and BAMRes18 models under different epochs; the middle column of Figure 6 indicates the Loss of the RGB/Flow stream of the Res50 and BAMRes50 models under different epochs; the third column of Figure 6 represents the Loss of the RGB/Flow stream of the Res101 and BAMRes101 models under different epochs, respectively.

Figure 5 and Figure 6 show that the accuracy and the loss of BAMRes18, BAMRes50, and BAMRes101 exceed that of Res18, Res50, and Res101, which validates the effectiveness of the bottleneck attention module.

These comparison results validate the effectiveness of the bottleneck attention unit. The reason was that bottleneck attention units allow the novel model to extract the spatial-channel dependencies on the appearance and optical flow streams, and can extract richer and more robust features, thus enhancing the performance of Jinnan cattle action recognition.

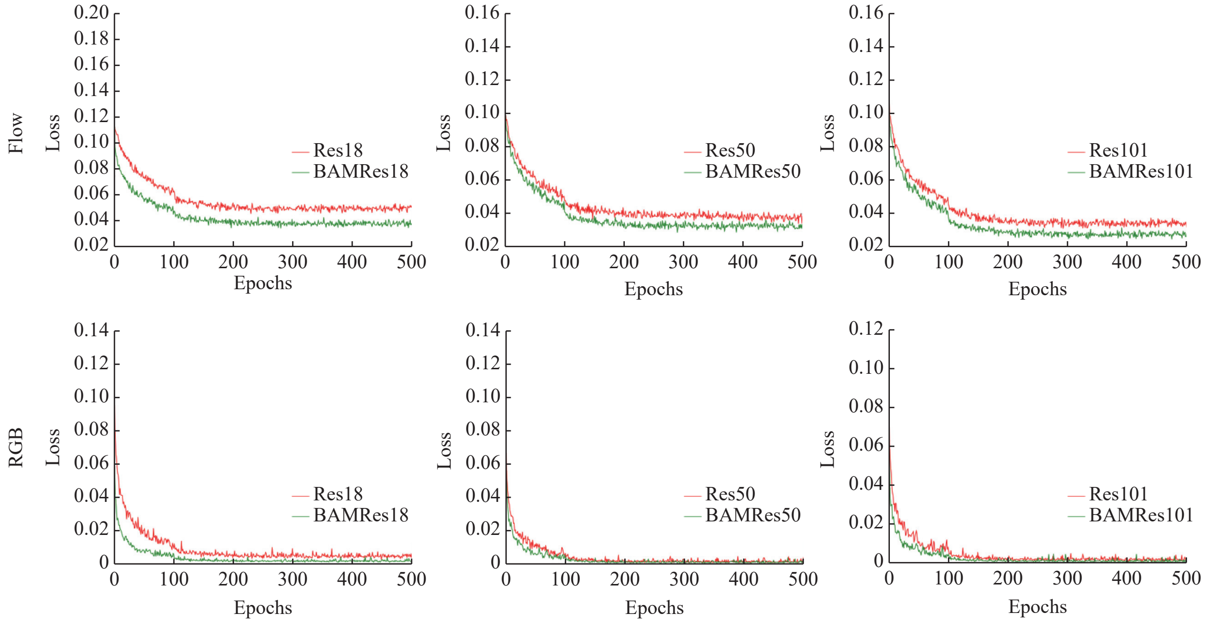
4 Conclusions

This study developed a novel Jinnan cattle action recognition method based on bottleneck attention enhanced two-stream neural network, which aims at exploring the spatial and channel dependencies between the RGB and flow stream of the video, for achieving superior action recognition performance. Furthermore, a new Jinnan cattle action recognition dataset which contains 1821 videos has been developed for prompting the intelligent breeding of Jinnan cattle. Experimental results validate the effectiveness of the proposed model BATS. Moreover, the BATS achieves state-of-the-art action recognition results, 4.6% better than the other models. Future work will focus on enriching the dataset with more diverse actions and environments to boost recognition robustness and accuracy. The authors' team also aim to integrate multimodal information, such as audio and temperature data, to further enhance the system's intelligence.



a. Flow b. RGB c. Fusion

Figure 5 Comparison results of different models with or without bottleneck attention module in terms of accuracy



a. ResNet18 b. ResNet50 c. ResNet101

Figure 6 Comparison results of different models with or without bottleneck attention module in terms of loss

Acknowledgements

This work was supported by the Shanxi Province Basic Research Program (Grant No. 202203021212444); Shanxi Agricultural University Science and Technology Innovation Enhancement Project (Grant No. CXGC2023045); Shanxi Province Higher Education Teaching Reform and Innovation Project (Grant No. J20220274); Shanxi Postgraduate Education and Teaching Reform Project Fund (2022YJJG094); Shanxi Agricultural University Doctoral Research Start-up Project (Grant No. 2021BQ88); Shanxi Agricultural University Academic Restoration Research Project (2020xshf38).

[References]

[1] Cairo F C, Pereira L G R, Campos M M, Tomich T R, Coelho S G, Lage C A, et al. Applying machine learning techniques on feeding behavior data for early estrus detection in dairy heifers. Computers and Electronics in Agriculture, 2020; 179: 105855.

[2] Tang J L, Yang G X, Sun Y R, Xin J, He D J. Salient object detection of dairy goats in farm image based on background and foreground priors. Neurocomputing, 2019; 332: 270–282.

[3] Shang C, Wu F, Wang M L, Gao Q. Cattle behavior recognition based on feature fusion under a dual attention mechanism. Journal of Visual Communication and Image Representation, 2022; 85: 103524.

[4] Heo E J, Ahn S J, Choi K S. Real-time cattle action recognition for estrus detection. KSII Transactions on Internet and Information Systems (TIIS), 2019; 13(4): 2148–2161.

[5] Nguyen C, Wang D, Von Richter K, Valencia P, Alvarenga F A P, Bishop-Hurley G. Video-based cattle identification and action recognition. In: 2021 Digital Image Computing: Techniques and Applications (DICTA), Gold Coast: IEEE, 2021; pp.1–5.

[6] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp.4724–4733.

[7] Yao G L, Tao L, Zhong J D. A review of convolutional-neural-network-based action recognition. Pattern Recognition Letters, 2019; 118: 14–22.

[8] Zhang H B, Zhang Y X, Zhong B N, Lei Q, Yang L J, Du J X, et al. A comprehensive survey of vision-based human action recognition methods. Sensors, 2019; 19(5): 1005.

[9] Ahn S J, Ko D M, Heo E J, Choi K S. Real-time cow action recognition based on motion history image feature. In: 2018 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas: IEEE, 2018; pp.1–2.

[10] Wu D H, Wu Q, Yin X Q, Jiang B, Wang H, He D J, et al. Lameness detection of dairy cows based on the YOLOv3 deep learning algorithm and a relative step size characteristic vector. Biosystems Engineering, 2020; 89: 150–163.

[11] Jiang B, Yin X Q, Song H B. Single-stream long-term optical flow convolution network for action recognition of lameness dairy cow. Computers and Electronics in Agriculture, 2020; 175: 105536.

[12] Shen W Z, Hu H Q, Dai B S, Wei X L, Sun J. Individual identification of dairy cows based on convolutional neural networks. Multimedia Tools and Applications, 2020; 79: 14711–14724.

[13] Tassinari P, Bovo M, Benni S, Franzoni S, Poggi M, Mammi L M E, et al. A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. Computers and Electronics in Agriculture, 2021; 182: 106030.

[14] Zhang X D, Kang X, Feng N N, Liu G. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. Computers and Electronics in Agriculture, 2020; 178: 105754.

[15] Hu H Q, Dai B S, Shen W Z, Wei X L, Sun J, Li R Z, et al. Cow identification based on fusion of deep parts features. Biosystems Engineering, 2020; 192: 245–256.

[16] Jiang B, Wu Q, Yin X Q, Wu D H, Song H B, He D J. FlyYOLOv3 deep learning for key parts of dairy cow body detection. Computers and Electronics in Agriculture, 2019; 166: 104982.

[17] Bezen R, Edan Y, Halachmi I. Computer vision system for measuring individual cow feed intake using RGB-D camera and deep learning algorithms. Computers and Electronics in Agriculture, 2020; 172: 105345.

[18] 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). Biosystems Engineering, 2020; 198: 31–49.

[19] Wu D H, Wang Y F, Han M X, Song L, Shang Y Y, Zhang X Y, et al. Using a CNN-LSTM for basic behaviors detection of a single dairy cow in a complex environment. Computers and Electronics in Agriculture, 2021; 182: 106016.

[20] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014; 1: 568–576.

[21] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Comuniacations of the ACM, 2017; 60(6): 84–90.

[22] Park J, Woo S, Lee J-Y, Kweon I. BAM: Bottleneck attention module. arXiv in Press, 2018. arXiv: 1807.06514.

[23] Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. MobileNetV2: Inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City: IEEE, 2018; pp.4510–4520.

[24] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv Preprint, 2014. arXiv: 1409.1556.

[25] Iandola F N, Han S, Moskewicz M W, Ashraf K, Dally W J, Keutzer K. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv Preprint, 2016. arXiv: 1602.07360.

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