Poultry Science

Behavior recognition of cage-free multi-broilers based on spatiotemporal feature learning

Yilei Hu,[*,†] Jiaqi Xiong,[*,†] Jinyang Xu,[*,†] Zhichao Gou,[*,†] Yibin Ying,[*,†] Jinming Pan,[*,†] and Di Cui[*,†,1]

[*]*College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, P. R. China; and* [†]*Key Laboratory of Intelligent Equipment and Robotics for Agriculture of Zhejiang Province, Hangzhou 310058, P. R. China*

ABSTRACT Poultry behavior indicates their health, welfare, and production performance. Timely access to broilers' behavioral information can improve their welfare and reduce disease spread. Most behaviors require a period of observation before they can be accurately judged. However, the existing approaches for multi-object behavior recognition were mostly developed based on a single-frame image and ignored the temporal features in videos, which led to misrecognition. This study proposed an end-to-end method for recognizing multiple simultaneous behavioral events of cage-free broilers in videos by Broiler Behavior Recognition System (**BBRS**) based on spatiotemporal feature learning. The BBRS consisted of 3 main components: the improved YOLOv8s detector, the Bytetrack tracker, and the 3D-ResNet50-TSAM model. The basic network YOLOv8s was improved with MPDIoU to identify multiple broilers in the same frame of videos. The Bytetrack tracker was used to track each identified broiler and acquire its image sequence of 32 continuous frames as input for the 3D-ResNet50-TSAM model. To accurately recognize behavior of each tracked broiler, the 3D-ResNet50-TSAM model integrated a temporal-spatial attention module for learning the spatiotemporal features from its image sequence and enhancing inference ability in the case of its image sequence less than 32 continuous frames due to its tracker ID switching. Each component of BBRS was trained and tested with the rearing density of 7 to 8 birds/m^2. The results demonstrated that the *mAP*@0.5 of the improved YOLOv8s detector was 99.50%. The Bytetrack tracker achieved a mean MOTA of 93.89% at different levels of occlusion. The *Accuracy*, *Precision*, *Recall*, and *F*1 *score* of the 3D-ResNet50-TSAM model were 97.84, 97.72, 97.65, and 97.68%, respectively. The BBRS showed satisfactory inference ability with an *Accuracy* of 93.98% when 26 continuous frames of the tracked broiler were received by the 3D-ResNet50-TSAM model. This study provides an efficient tool for automatically and accurately recognizing behaviors of cage-free multi-broilers in videos. The code will be released on GitHub (https://github.com/CoderYLH/BBRS) as soon as the study is published.

Key words: broiler, behavior recognition, spatiotemporal feature, end-to-end, computer vision

2024 Poultry Science 103:104314
https://doi.org/10.1016/j.psj.2024.104314

INTRODUCTION

China is the largest producer and consumer of poultry meat in the world, with poultry meat consumption accounting for about 25% of all meat consumption (China Animal Agriculture Association, 2022). Broiler production is an important part of poultry production. The natural behavioral expressions of broilers can reflect their health, welfare, and production performance (Yang et al., 2023a). Timely acquisition of behavioral information of broilers can improve their welfare and reduce the spread of diseases (Xiao et al., 2019). Currently, the broiler behaviors are inspected by farm workers, who identify health abnormalities of broilers based on their behavioral expressions. However, this method is labor-intensive, subjective, and inefficient (He et al., 2022). With an aging population and rising labor costs, an automated and intelligent method is needed to replace the manual inspection to meet the development needs of modern large-scale poultry farming enterprises (Zhao, 2019).

Computer vision, as an emerging non-destructive detection technology, has the advantages of being efficient, non-stress inducing, and low-cost. It is regarded as an effective means for monitoring poultry behaviors (Handan-Nader and Ho, 2019). The existing approaches for multi-object behavior recognition were mostly

Received June 18, 2024.
Accepted September 5, 2024.

[1]Corresponding author. dicui@zju.edu.cn

developed based on a single-frame image. The object detectors such as the You Only Look Once (**YOLO**) and the Region-Based Convolutional Neural Network (**R-CNN**) were usually applied to recognize poultry behaviors including resting, feeding, drinking, pecking at feathers, standing, fighting, exploring, and mating from the single-frame image (Wang et al., 2020; Subedi et al., 2023; Yang et al., 2023b). To enhance the robustness of the detectors, a combination of techniques including data augmentation, attention mechanisms, improved loss functions, and improved feature fusion strategies were deployed (Liu et al., 2023; Xiao et al., 2023). However, most behaviors usually require a period of observation before they can be accurately judged. The methods based on a single-frame image ignored the temporal features of behaviors and suffered from semantic ambiguities due to high spatial feature similarities between different behaviors, which easily lead to misrecognition.

To effectively utilize the temporal features of poultry behaviors, a feature learning model was used to extract spatial features from the continuous single-frame images first, followed by employing a sequential model such as Long Short-Term Memory Network (**LSTM**) to learn the temporal dependencies of the spatial features (Fang et al., 2021; Nasiri et al., 2022; Volkmann et al., 2022). Alternatively, a video understanding model such as the Temporal Shift Module (**TSM**) was utilized to simultaneously extract both the spatial and temporal features from videos, which was reported in the field of livestock behavior recognition and held significant potential for application in poultry behavior recognition (Ji et al., 2023). However, these methods involved clipping a continuous image sequence of a single behavioral event from the raw video, or ensuring that the raw video contained only one behavioral event. In a word, they were primarily designed for scenarios with single behavioral events, making their direct application challenging due to the simultaneous occurrence of multiple behavioral events.

The general objective of this study is to develop an end-to-end method for recognizing multiple simultaneous behavioral events of cage-free broilers in videos. The specific objectives are: 1) to improve the the basic YOLOv8s detector for identifying multiple broilers in the same frame of videos, 2) to employ an algorithm for tracking each identified broiler to acquire its image sequence of 32 continuous frames, 3) to establish a model for recognizing broiler behaviors with the obtained image sequences as inputs, and 4) to evaluate the performance of the whole system consisted of the above 3 components.

MATERIALS AND METHODS

Materials

Experimental setup and Data acquisition All procedures of this experiment were performed under the guidance of the Care and Use of Animals of the Zhejiang University (Hangzhou, China). The Committee on the Ethics of Animal Experiments of Zhejiang University approved the protocol. As shown in Figure 1, 2 batches of experiments were conducted in a 3.2 m x 3.0 m x 2.6 m (lengthxwidthxheight) room equipped with a heating plate, air conditioner, and temperature and humidity sensors. The ground was filled with 5 cm thick litter. The light was on from 6:00 to 22:00, with an intensity of 25 lux. The camera (DS-2TD2636B-10/P (B), Hikvision, Hangzhou, China) was mounted at the center of the room at a height of 1.8 m from the ground, with a resolution of 2,688 × 1,520 pixels and a sampling frame rate of 25 fps. The central area, located within the camera's field of view (**FOV**), was 1.2 m in length and 1.1 m in width, and contained a drinker, feeder, and weight scale. To collect different behaviors of broilers, ten 18-day-old broilers (Youhuang 5B, male) were purchased in 2 batches from Jiangsu Lihua Animal Husbandry Co., Ltd. The broilers were acclimated in the central area for 2 d to familiarize themselves with the experimental environment. For each batch, the experiment began when the birds reached 20-days-old and ended at 33-days-old, lasting a total of 14 d. The temperature was maintained at 33°C from 9:00 to 18:00 and set to 23°C from 18:00 to 9:00 the next day for the first batch as experimental group, while the temperature was kept constant at 23°C for the second batch as control group. The videos of broilers were recorded from 9:00 to 18:00. Since the camera's FOV extended beyond the central area, the region of interest (ROI) in videos were limited to 1628×1520 pixels for eliminating irrelevant background regions. The ROI in videos were then utilized for subsequent video processing tasks.

Dataset Description In this study, 918 GB of videos were collected in two batches of experiments. Two datasets were derived from the videos for training and testing the components within the BBRS: an object recognition dataset and a behavior recognition dataset. Both datasets were randomly split into training, validation, and testing sets in a ratio of 7:1:2, as shown in Table 1. A

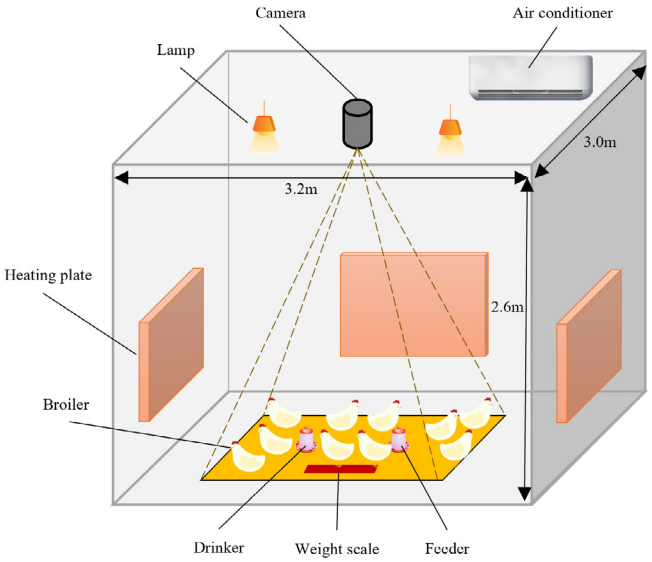


Figure 1. Diagram of the experimental system for collecting broiler behavior data.

Table 1. Dataset composition of object recognition and behavior recognition.

Dataset types	Total	Train	Validation	Test	Classes
Object recognition dataset	3,048	2,133	304	611	1
Behavior recognition dataset	7,153	5,005	709	1,439	5

Table 2. Definitions of broiler behaviors.

Types	Description	Number of video segments
Activity	Including walking and standing actions	1,556
Spreading wing	Spreading wings to expose the skin underneath	1,557
Resting	Lying on the bedding without the behavior of spreading wing	1,172
Feeding	The broiler head is positioned on the side of the feeder and the beak pecks continuously at the feed	1,525
Drinking	The broiler head is positioned on the side of the drinker and the beak repeatedly performs the actions of lowering the head to collect water and lifting the head to drink water	1,343

total of 3,048 images with 1 category of objects were manually selected from the videos. These images included the information on the broilers from both batches over the entire experimental period, covering the age range from 20 d to 33 d. They were labelled by LabelMe 5.3.1 (https://github.com/labelmeai/labelme) as the object recognition dataset for training and testing the improved YOLOv8s detector (Section *Improved YOLOv8s detector*). Additionally, a total of 7153 video segments with 5 categories of behaviors were extracted from the videos using the ByteTrack tracker (Section *Bytetrack tracker*). These video segments also covered the entire experimental period for both batches of broilers, capturing the age range from 20 d to 33 d. They served as the behavior recognition dataset for training and testing the 3D-ResNet50-TSAM model (section *D-ResNet50-TSAM behavior recognition model*). As shown in Table 2, the behavior categories included activity, spreading wing, resting, feeding, and drinking, with video segments numbers of 1,556, 1,557, 1,172, 1,525, and 1,343, respectively. To maintain consistency within behavior categories and discriminability between different behavior categories, and to avoid potential semantic ambiguities in category definitions, the detailed definitions of the 5 behavior types were also provided in Table 2.

Algorithm Development Environment and Software The algorithm development and testing platform used in this study is a computer equipped with an NVIDIA GeForce RTX 3090 GPU with 12GB of VRAM, a 13th Gen Intel(R) Core(TM) i5-13400 CPU, and runs on the Ubuntu 22.04 operating system. The project environment includes Python 3.10, OpenCV 4.8.1, PyTorch 2.0.0, torchvision 0.15.1, and CUDA 11.7.

Methods

Overall architecture of the BBRS To develop the method for recognizing multiple simultaneous behavioral events of cage-free broilers in videos, an end-to-end system named the Broiler Behavior Recognition System (**BBRS**) was proposed in this study, and its overall architecture was illustrated in Figure 2. The BBRS consists of 3 components: the improved YOLOv8s detector, the Bytetrack tracker, and the 3D-ResNet50-TSAM model. Initially, the improved YOLOv8s detector is used to perform frame-by-frame detection on the original video and output bounding boxes of the detected broilers. Subsequently, the Bytetrack tracker is employed to construct tracking trajectories and assign tracking IDs for each individual broiler across different frames. Based on the tracking ID and bounding box of each broiler, a continuous image sequence of 32 frames is extracted. The obtained continuous image sequences of all the broilers are then fed into the 3D-ResNet50-TSAM behavior recognition model, which performs synchronous inference to classify the corresponding behavior category for each broiler.

The key module of the BBRS is the tracking result processing algorithm, which obtains the tracking results and utilizes them for 3D-ResNet50-TSAM model inference. The detailed steps of this algorithm are outlined below, and its pseudo-code is represented in Algorithm 1.

1. The tracking results list includes broiler IDs, bounding boxes, frame IDs, and confidence scores, and is assumed to maintain a continuous image sequence of length T for each broiler.
2. For each broiler within a time window of length T:
    a. If the tracking IDs are continuous, extract the corresponding continuous T-frame bounding boxes for each ID, compute the maximum bounding rectangle for the T-frame bounding boxes, and derive the corresponding image sequence from the original video using this maximum bounding rectangle. This results in n sets of image sequences of the size of the maximum bounding rectangle and length of T (where n is the number of broilers).
    b. For broilers whose k tracking IDs are not continuous ($k < n$), extract bounding boxes for m frames ($m < T$), where m is the last frame before the ID switches, calculate the maximum bounding rectangle for these frames, and derive the required m-frame image sequence from the original video using this bounding rectangle. The deficient T-m frames are padded with blank frames of the same size. The image sequences for the remaining n-k broilers are procured as described in part a., ensuring a

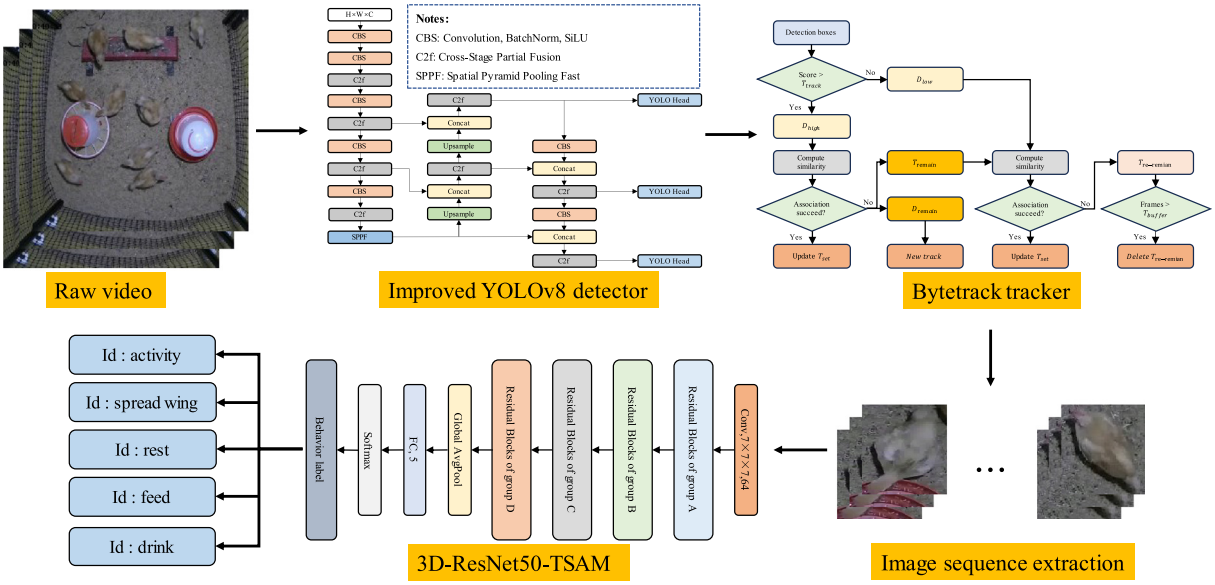


Figure 2. Overall architecture of the BBRS for broiler behavior recognition.

total of n sets of image sequences of the size of the maximum bounding rectangle and length of T.

3. Rescale the n sets of image sequences to a uniform size of 224 × 224, yielding data with the form (B_0, T, C, H, W), where B_0 is the batch size equal to the number of broilers n, T is the length of a time window, C is the number of channels, set to 3, and H and W are the height and width of the input, both with values of 224.
4. Feed the data shaped (B_0, T, C, H, W) into the 3D-ResNet50-TSAM model for inference, achieving the behavioral labels corresponding to each broiler ID.

Improved YOLOv8s Detector To realize the recognization of broiler behaviors, the first step was to identify all the broilers in each frame. Therefore, the basic YOLOv8s detector was improved to do this task. The improved YOLOv8s detector mainly comprises of the Backbone, the Neck, and the Head (Figure S1). Initially, images are fed into the Backbone to extract feature maps, which then pass through the Neck for bidirectional feature fusion to enhance multi-scale representation capability of the extracted features. Ultimately, decoupling operations are performed in the Head to separately utilize the feature maps for classification and regression tasks.

The basic YOLOv8s detector uses the Complete IoU (**CIoU**) Loss as the regression loss to compute the overlap area, center point distance, and aspect ratio between the predicted and ground-truth boxes (Zheng et al., 2020). However, it has shown to be less accurate and slower in convergence when the predicted and ground-truth boxes share the same aspect ratio but differ in size. To overcome the limitation, the improved YOLOv8s detector replaces CIoU Loss with Minimum Point Distance based IoU (**MPDIoU**) Loss. MPDIoU Loss minimizes the Euclidean distance between the 4 vertices of the predicted and ground-truth boxes, calculated as eqns. (1 and 2).

$$MPDIoU = IoU - \frac{\left(x_1^{gt} - x_1^{pred}\right)^2 + \left(y_1^{gt} - y_1^{pred}\right)^2}{w^2 + h^2} - \frac{\left(x_2^{gt} - x_2^{pred}\right)^2 + \left(y_2^{gt} - y_2^{pred}\right)^2}{w^2 + h^2} \quad (1)$$

$$L_{MPDIoU} = 1 - MPDIoU \quad (2)$$

where IoU is the Intersection over Union of the predicted and ground-truth boxes, (x_1^{gt}, y_1^{gt}) and (x_2^{gt}, y_2^{gt}) are the coordinates of the top-left and bottom-right corners of the ground-truth box, respectively, and (x_1^{pred}, y_1^{pred}) and (x_2^{pred}, y_2^{pred}) are the coordinates of the top-left and bottom-right corners of the predicted box, respectively,

Algorithm 1. Tracking result processing algorithm for behavior recognition.

Input: List of tracking results L, time window length T, video V.
Output: Behavioral labels O for each target ID.

Initialize image sequences array $S[n]$ for n targets
for ($i = 1$ to n) **do**
  B_box = extract_bounding_boxes($L[i]$, T);
  if (is_continuous(B_box)) **then**
    R = compute_max_rectangle(B_box); //Compute the maximum bounding rectangle
    $S[i]$ = derive_sequence(V, R, T); //Derive the image sequence
  Else
    m = find_switch_frame(B_box, T); //Find the last frame before the ID switches
    R = compute_max_rectangle($B\_box[1{:}m]$);
    $S[i]$ = derive_sequence(V, R, m);
    $S[i]$ = pad_sequence($S[i]$, T); //Blank frames padding
  end if
end for
for ($i = 1$ to n) **do**
  $S[i]$ = rescale_sequence($S[i]$, 224, 224); //Normalization
end for
O = model_inference(S); //Behavior recognition
Update tracking results L for model inference based on next time window length T;
return O;

with w and h being the width and height of the input image (Ma and Xu, 2023).

By applying MPDIoU Loss, the improved YOLOv8s detector achieves higher sensitivity in localizing broiler targets of varying sizes and postures, particularly in dealing with individuals with partial occlusions where this novel loss function provides more accurate and stable bounding boxes, delivering high-quality observational data to the BBRS's tracker.

The improved YOLOv8s detector was trained using the object recognition dataset for 150 epochs, with a batch size of 12, and selecting the Stochastic Gradient Descent (**SGD**) optimizer, while the rest of the training parameters remained default. The performance of the improved YOLOv8s detector was evaluated on the test set.

Bytetrack Tracker After the broilers were identified, it is necessary to track them for acquiring continuous image sequences of their behavior expression process. The image sequences contains spatial and temporal information which helps to make more accurate judgments about broiler behaviors. Most trackers typically associate targets between consecutive frames by comparing similarities in their position, appearance, and motion characteristics to generate tracking trajectories (Rakai et al., 2022). However, in the field of poultry, the high similarity in the appearance features of broilers leads to difficulties in achieving performance advantages with appearance-based trackers, which in turn increases computational complexity. Consequently, this study opts for the Bytetrack tracker that does not rely on an appearance feature branch.

The principle of the Bytetrack tracker entails conducting object detection on images to acquire bounding boxes with associated confidence scores (Figure S2). The obtained boxes are then classified into high-score boxes D_{high} and low-score boxes D_{low} based on confidence scores threshold T_{track}. The kalman filtering is then used to predict and update the status of existing trajectory sets T_{set}. IoU is computed between D_{high} and T_{set}, after which the Hungarian matching algorithm is employed for the first association between T_{set} and D_{high}. The sets of unmatched detection boxes and trajectories are respectively denoted as D_{remain} and T_{remain}. The unmatched high-score detection boxes in D_{remain} are initialized as new trajectories into the trajectory sets T_{set}. Subsequently, the Hungarian matching algorithm is used for a second association between the remaining trajectories T_{remain} and low-score boxes D_{low}. Each trajectory in T_{remain} records the number of consecutive unmatched frames, and if this number exceeds a predefined buffer threshold T_{buffer}, the trajectory is deleted (Zhang et al., 2022). This mechanism enables the Bytetrack tracker to provide robust performance in high-density poultry farming scenes.

The Bytetrack tracker utilizes 3 hyperparameters: T_{buffer}, T_{track}, and T_{match}. T_{buffer} is set to 30, determining the maximum number of frames a trajectory can retain in tracking status before being marked as lost. T_{track} is set to 0.5, which serves to differentiate between high-score and low-score detection boxes. T_{match} is the IoU threshold between the predicted and detection boxes, set to 0.8.

The stability of the Bytetrack tracker is a crucial prerequisite for the BBRS to receive a continuous image sequence of target broiler. It determines the subsequent accuracy of behavior recognition. In experiments, the stability of the tracker is influenced by factors such as broilers occlusion and motion blur. Compared to transient interference caused by sudden motion blur, dynamic occlusion of broilers is a continuous and critical factor affecting tracker's performance. This research defines the average occlusion degree OD_{avg} of the video to quantify the impact of broiler occlusion events on tracker performance and standard deviation $\sigma_{OD_{avg}}$ to describe the variability of occlusion degree throughout the video, calculated as Eqs. (3)-(5).

$$OD_{frame} = \left(1 - \frac{B_{frame}}{B_{complete}}\right) \times 100\% \tag{3}$$

$$OD_{avg} = \frac{1}{N}\sum_{i=1}^{N} OD_{frame_i} \tag{4}$$

$$\sigma_{OD_{avg}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(OD_{frame_i} - OD_{avg}\right)^2} \tag{5}$$

Where OD_{frame} is the occlusion degree of a single frame, OD_{avg} is the average occlusion degree of a video, $\sigma_{OD_{avg}}$ is the standard deviation of occlusion degree throughout the video, N is the number of frames in a video, B_{frame} is the number of broiler pixels in the image during an occlusion event, and $B_{complete}$ represents the number of broiler pixels in the image when there is no occlusion event. B_{frame} and $B_{complete}$ are calculated as follows: each frame of the test video is frame-by-frame detected using the improved YOLOv8s detector to obtain each frame's bounding boxes, which serve as prompts for the Segment Anything Model (**SAM**) to perform semantic segmentation and obtain broiler masks (Kirillov et al., 2023). For masks with substandard segmentation quality, a new prompt for the SAM model involving a combination of bounding boxes from the improved YOLOv8s detector and points provided by the operator is utilized to predict new masks, ensuring improved segmentation accuracy through this tailored approach. The number of mask pixels for each frame is then tallied.

3D-ResNet50-TSAM Behavior Recognition Model To extract the spatial and temporal features hidden in the obtained image sequences, a behavior recognition model 3D-ResNet50-TSAM was established. The 3D-ResNet50-TSAM model is built upon the basic 3D-ResNet network by integrating a Temporal-Spatial Attention Module (**TSAM**), which enables the model to autonomously learn important regions and key frames from the obtained image sequences. The 3D-ResNet50-TSAM model processes a continuous image sequence of

broilers through 3D convolution layers and residual blocks of groups A, B, C, and D for feature extraction (Figure S3). The features' spatial dimensions are then compressed through a global average pooling layer. Subsequently, a Fully Connected (**FC**) layer models the non-linear relationships of the features, and finally, a softmax function maps the output to probability values to obtain the behavior recognition results (He et al., 2016; Hara et al., 2018). The number of residual blocks within groups A, B, C, and D are 3, 4, 6, and 3, respectively, and the last residual block in each group is an improved residual block (Figure S4a), while the rest are basic residual blocks (Figure S4b). Replacing only the last residual block of each group with the improved residual block, as opposed to all of them, not only enhances the model's ability to learn deeper features but also reduces the computational burden and prevents overfitting.

Two main improvements are applied to the basic residual block to form the improved residual block. Firstly, the ReLU activation function in the basic residual block has been replaced with the Mish activation function, defined in eqns. (6 and 7).

$$f(x) = x \cdot \tanh(\varsigma(x)) \tag{6}$$

$$\varsigma(x) = \ln(1 + e^{x}) \tag{7}$$

Where $\varsigma(x)$ is the softplus function.

The Mish activation function is smooth, continuous, and non-monotonically increasing. Compared to the ReLU activation function used in the basic residual block, the Mish activation function, owing to its non-saturating nature, can mitigate the vanishing gradient problem in deep networks. Moreover, due to its continuity and differentiability, it maintains an effective gradient flow during backpropagation, which aids in efficient weight updates during training, enhancing the adaptation of the model to the distribution of activation values (Misra, 2020). This improvement bolsters the learning representation capability and generalization of the 3D-ResNet50-TSAM model.

Secondly, at the end of the last batch normalization (BN) layer in the basic residual block, the TSAM has been appended. TSAM is a fusion of temporal and spatial attention mechanisms that provide several advantages over current popular spatio-temporal attention mechanisms based on self-attention, such as the Non-Local (**NL**) block(Wang et al., 2018). These advantages include lower computational complexity, higher efficiency, and better interpretability. The operation of TSAM is as follows: Given an input feature map sequence $\{T_1, T_2 \cdots T_n\}$, where n is the number of feature maps, and each feature map $T_i (1 \leq i \leq n)$ is of size H×W×C, with H, W, and C corresponding to the height, width, and number of channels of the feature map, respectively. Each feature map T_i is used to perform max pooling and average pooling operations along the channel direction, resulting in max-pooled feature maps $W_{i-1-max}$ and average-pooled feature maps $W_{i-1-avg}$ both of size H×W×1. The feature maps $W_{i-1-max}$ and $W_{i-1-avg}$ are element-wise summed to yield a feature map W_{i-1} of size H×W×1. Global max pooling and global average pooling operations are subsequently applied to the feature map W_{i-1}, producing max-pooled feature maps $W_{i-2-max}$ and average-pooled feature maps $W_{i-2-avg}$, each of size 1×1×1. Adding $W_{i-2-max}$ and $W_{i-2-avg}$ feature maps together, a feature map W_{i-2} of size 1×1×1 is obtained. After passing the feature map sequence W_{i-2} through the sigmoid activation function, temporal weights W'_{i-2} are obtained, which, when multiplied with the feature map T_i, result in a weighted feature map sequence $\{T'_1, T'_2 \cdots T'_n\}$ of size H×W×C. This operation emphasizes the contribution of important temporal frames while suppressing interference from irrelevant ones. Building on this, each weighted feature map T'_i from the sequence $\{T'_1, T'_2 \cdots T'_n\}$ undergoes parallel max pooling and average pooling operations along the channel direction, leading to max-pooled maps $W_{i-3-max}$ and average-pooled maps $W_{i-3-avg}$, both of size H×W×1. The addition of feature maps $W_{i-3-max}$ and $W_{i-3-avg}$ results in a feature map W_{i-3} of size H×W×1. After applying the sigmoid activation function, spatial weights W'_{i-3} are obtained and, when multiplied with T'_i, produce the final output feature map sequence $\{T''_1, T''_2 \cdots T''_n\}$ of size H×W×C. This operation is utilized to emphasize learning of important regions within each temporal frame.

Moreover, due to disparities in the number of video segments for the 5 broiler behaviors in the behavior recognition dataset, an effective solution to address the class imbalance problem and enhance model performance for predicting behaviors with fewer samples is employed during the training of the 3D-ResNet50-TSAM model. This solution involves the use of a weighted cross-entropy loss function, L_{wce}. L_{wce} modifies the traditional cross-entropy loss function by introducing corresponding weight coefficients for each behavior class, thereby adjusting the impact of each class on the overall loss. The formula for L_{wce} is as follows in eqn. (8).

$$L_{wce} = -\sum_{i=1}^{N} (w_i y_i \log(\hat{y}_i)) \tag{8}$$

where N represents the number of behavior classes, w_i is the weight coefficient for the i-th behavior class, y_i is the true label for the i-th behavior, and $\hat{y}_i$ is the predicted probability for the i-th behavior by the model.

For the training of the 3D-ResNet50-TSAM model, a series of data augmentation techniques are employed to enhance the model's generalization capability in behavior recognition tasks. These techniques include adjustments in contrast, brightness, saturation, as well as horizontal flipping, vertical flipping, and random cropping. The network weights and parameters are updated using SGD optimizer, with the weighted cross-entropy loss function L_{wce} quantifying the discrepancies between model predictions and true values. Key hyperparameters that require fine-tuning during the model training process encompass the training epoch, batch size, input

Table 3. Parameter settings for training the 3D-ResNet50-TSAM model.

Parameters	Value
Training epoch	150
Batch size	4
Input size	224
Learning rate	0.01
Momentum	0.9
Weight decay	0.0001
Sample duration	32
Sample stride	1

size, learning rate, momentum, and weight decay, as well as the sample duration and sample stride, which are essential for temporal data augmentation. The training parameters for the 3D-ResNet50-TSAM model are presented in Table 3.

In the testing phase, to comprehensively assess the performance of the fully trained model across different temporal segments of the video, a temporal data augmentation strategy is employed, where subsequences are extracted from each original 64-frame sequence at distinct intervals (frames 0-31, 16-47, and 32-63).

Performance Evaluation Metrics for the BBRS

This study evaluated not only the performance of each component of BBRS, but also that of the whole BBRS.

For the improved YOLOv8s detector, the performance evaluation metrics include mean Average Precision (mAP), *Precision*, *Recall*, and $F1$ *score*, calculated by eqns. (9-12).

$$mAP = \frac{\sum_{i=1}^{n} AP(i)}{n} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{12}$$

Where TP and FP are the number of correctly and incorrectly detected broilers, respectively, and FN is the number of broilers missed for detection.

For the Bytetrack tracker, the performance evaluation metrics include Multiple Object Tracking Accuracy ($MOTA$), $IDF1$ *score*, and *ID* Switches ($IDSW$), calculated by eqns. (13 and 14).

$$MOTA = 1 - \frac{FN + FP + IDSW}{GT} \tag{13}$$

$$IDF1\ score = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{14}$$

Where $IDSW$ is the number of *ID* switches. $IDTP$, $IDFP$, and $IDFN$ respectively stand for the numbers of true positive *IDs*, false positive *IDs*, and false negative *IDs*, while GT is the number of the ground truth bounding boxes.

For the 3D-ResNet50-TSAM behavior recognition model, the employed performance metrics are *Precision*, *Recall*, $F1$ *score*, and *Accuracy*, calculated by eqns. (10-12) and (15).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

Where TP and FP are the number of correctly and incorrectly recognized behaviors, respectively, FN is the number of behaviors missed for recognition, and TN is the number of correctly identified absence of behaviors.

RESULTS AND DISCUSSION

Performance of the Improved YOLOv8s Detector

Table 4 presents the performance of the YOLOv8s detector before and after improvements. Prior to the improvement, despite interference factors such as mutual occlusion among broilers, incomplete broiler presentation, dim lighting, and low contrast between foreground and background, the basic YOLOv8s detector still achieved commendable performance on the test set. The values for $F1$ *score*, *Precision*, *Recall*, and mAP@0.5 were 99.58%, 99.54%, 99.62%, and 99.29%, respectively. These figures indicated that the object recognition task in this study was not overly difficult, which was highly associated with the number of broilers per unit area as well as the size of the broilers in the image. Building on this foundation, the improved YOLOv8s detector yielded a 0.21% increase in mAP@0.5, providing the BBRS's tracker with stable and reliable observations.

Performance of the Bytetrack Tracker

In this study, occlusion levels were classified into 3 categories: high, medium, and low, with corresponding OD_{avg} values ranging between 0-10%, 10-20%, and 20-30%, respectively. Two 30-second video segments from each occlusion category were selected to test the performance of the ByteTrack tracker.

The test results of the ByteTrack tracker are shown in Table 5. The mean $MOTA$, $IDF1$ *score*, and $IDSW$ under the 3 different occlusion levels were 93.89%, 87.55%, and 5, respectively. For low occlusion scenarios, where broilers were relatively scattered, the $MOTA$ could reach over 98%, with the $IDF1$ *score* of around 93%. This indicated that the ByteTrack tracker achieved accurate and stable tracking of broilers under low occlusion conditions. Medium occlusion scenarios presented a decline in tracking performance due to the tracker having to handle random occlusions caused by

Table 4. Performance of improved YOLOv8s detector.

Model	*F*1 *score* (%)	*Precison* (%)	*Recall* (%)	*mAP*@0.5 (%)
YOLOv8s	99.58	99.54	99.62	99.29
YOLOv8s + MPDIoU	99.77	99.73	99.81	99.50

Table 5. Performance of Bytetrack tracker under different occlusion levels.

Occlusion level	Video sequence	OD_{avg} (%)	$\sigma_{OD_{avg}}$ (%)	*MOTA* (%)	*IDF*1 *score* (%)	*IDSW*
Low	seq-1	5.94	2.13	98.77	93.95	2
	seq-2	7.04	3.34	98.19	93.22	3
Medium	seq-3	14.34	3.23	94.45	88.41	4
	seq-4	16.42	3.91	94.89	87.85	4
High	seq-5	24.86	3.19	88.67	81.24	8
	seq-6	25.43	3.55	88.36	80.64	9

the movement of broilers. For high occlusion scenarios, the *MOTA* was below 90%, the *IDF*1 *score* was around 80%, and the *IDSW* was 8. Comparatively, the *MOTA* dropped by about 10%, *IDF*1 *score* decreased by approximately 13.5%, and the number of *IDSW* increased 3-fold from the low occlusion scenarios. This demonstrated that the tracker faced significant challenges in stable association of multiple broilers with severe mutual occlusion.

The experiment revealed the strengths and limitations of the ByteTrack tracker in handling scenarios with varying degrees of broiler occlusion. The tracking performance of the ByteTrack tracker decreased with ascending occlusion levels, and the number of *IDSW* increased accordingly. Additionally, in the BBRS pipeline, every 32-frame continuous image sequence of each broiler is fed into the 3D-ResNet50-TSAM model for inference to obtain behavior recognition results. During this 32-frame image sequence, if an ID switching occurs, the 3D-ResNet50-TSAM model will not receive a complete and valid image sequence. The images from the frame where the ID switching occurs to the 32nd frame will be padded with blank frames, while only the images from the first frame to the last frame before the ID switching contain valid information of the broiler. Therefore, an ID switching occurring within this 32-frame sequence directly impacts the amount of valid information the 3D-ResNet50-TSAM model can receive and process, thereby affecting the accuracy of behavior recognition.

Performance of the 3D-ResNet50-TSAM model

Figure 3 depicts the confusion matrix of the 3D-ResNet50-TSAM on the test dataset. The model exhibited *Accuracy*, *Precision*, *Recall*, and *F*1 *score* of 97.84, 97.72, 97.65, and 97.68%, respectively. The results indicated the high accuracy and stability of the behavior recognition method. For specific behaviors, the *Precision* for activity, spreading wing, resting, feeding, and drinking were 98.07, 96.00, 95.18, 99.61, and 99.73%, respectively. Feeding and drinking behaviors had the highest *Precision*, followed by activity, with spreading wing and resting having relatively lower *Precision*. There were 15 samples in which the behavior of resting was incorrectly identified as spreading wing, possibly because the broilers, while resting, exhibited small degrees of spreading wing that led the model to misinterpret the behavior as spreading wing. Moreover, 3 samples of spreading wing behavior were misconstrued as activity, likely due to the presence of larger movements in conjunction with spreading wing, causing the model to classify it as activity. Additionally, 3 samples of spreading wing behavior were misclassified as resting because the degree of spreading wing was small, and the broilers were almost motionless at that time window. This indicates that when the behavior characteristics of broilers are not sufficiently distinct, there is a reduced differentiation between behaviors, directly affecting the model's recognition capability. To address these issues, prior expert knowledge on the behavioral characteristics of resting and spreading wing can be integrated into the BBRS. Extensive experience in recognizing broiler behaviors is held by ethologists. The experience can be transformed into text and incorporated into the model, making the BBRS a mechanism and data-driven system, thereby

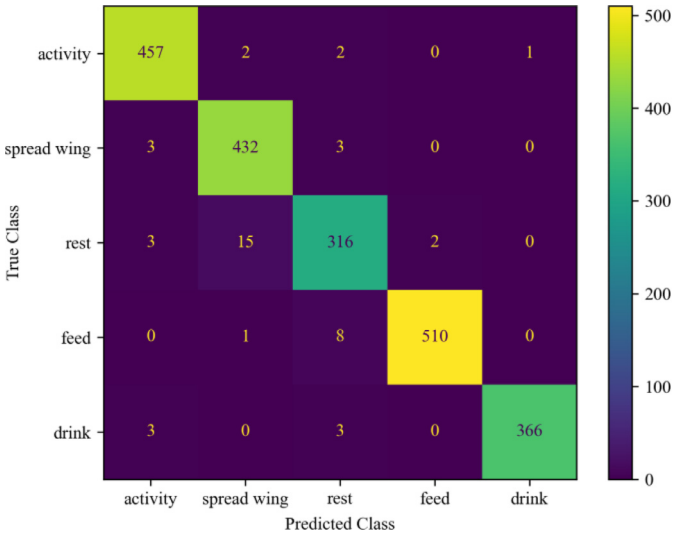


Figure 3. The confusion matrix of the 3D-ResNet50-TSAM on the test dataset.

Table 6. Contribution of different modules to the 3D-ResNet50-TSAM model.

Experiment number	TSAM	Mish	L_{wce}	*Accuracy* (%)
1	×	×	×	95.30
2	×	×	√	95.71
3	×	√	×	95.66
4	√	×	×	97.45
5	√	√	√	97.84

enhancing its accuracy. Furthermore, the feature extraction capability of the BBRS can be improved by finer-grained motion analysis to extract more effective features for better distinguishing between subtle variations in resting and spreading wing behavior.

To further investigate the impact of individual modules on the performance of the 3D-ResNet50-TSAM model, a series of ablation studies were conducted, with the results shown in Table 6. Upon replacing all ReLU activation functions with Mish activation functions in the basic residual blocks of the 3D-ResNet50 model, there was an observed increase in *Accuracy* of 0.36%. When substituting the cross-entropy loss function with the L_{wce} during the training of the 3D-ResNet50 model, the *Accuracy* improved by 0.41%. The integration of the TSAM within the residual blocks of the 3D-ResNet50 model contributed to the *Accuracy* improvement of 2.15%, indicating a significant role of TSAM in boosting the model's performance. The comprehensive adoption of TSAM, Mish activation, and L_{wce} in the final 3D-ResNet50-TSAM model, manifested in a cumulative accuracy enhancement of 2.54%. This outcome suggests that the combination of the aforementioned 3 modules effectively and significantly contributes to the performance improvement of the 3D-ResNet50-TSAM model for behavior recognition.

To delve deeper into the operational mechanisms of the 3D-ResNet50-TSAM model during inference, this study separately analyzes the model's dependence on temporal and spatial attention for decision-making. Figure 4 presents the temporal attention weight curves for 5 broiler behaviors. Observing the distribution of attention weights for these behaviors over various frames revealed a differentiated focus trend during model inference. For behaviors such as resting, activity, and spreading wing, the model tended to pay more attention to the first 8 frames of the image sequences, assigning equal weight to these frames. This implied that, in the model's cognition, these continuous 8 frames were more important for model inference and held the same informational value. Conversely, for feeding and drinking behaviors, the model concentrated on the last 8 frames of the image sequences, indicated that these final frames played a crucial role in the recognition of feeding and drinking behaviors.

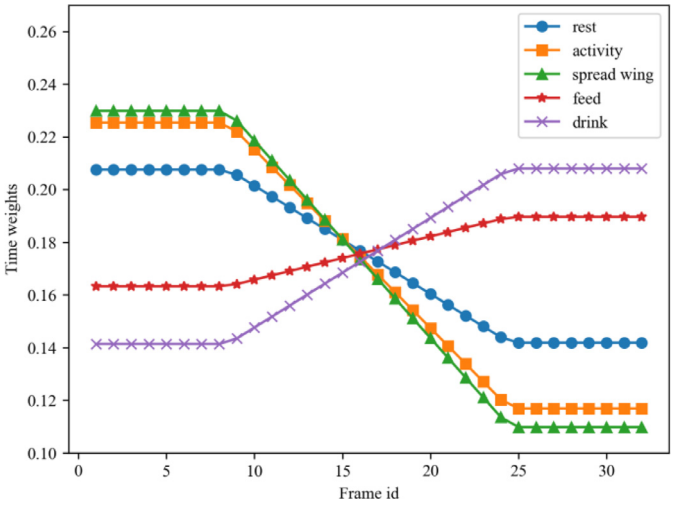
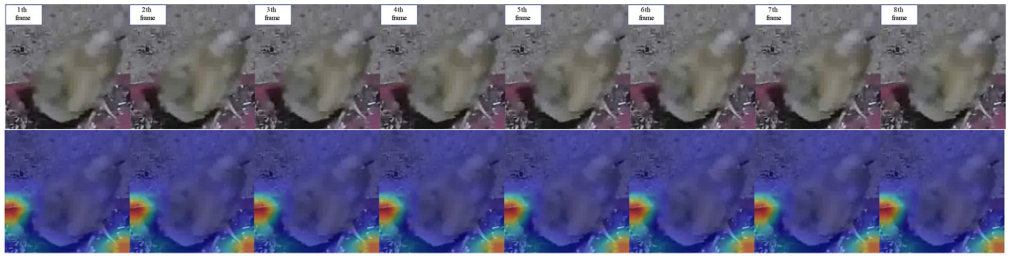


Figure 4. Temporal attention weight curves of 5 broiler behaviors.

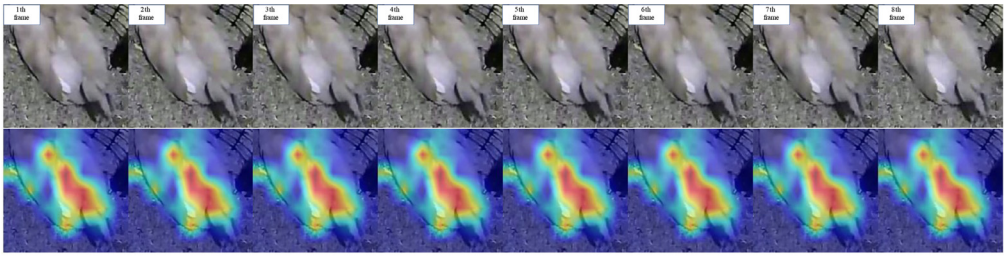
Additionally, it was noteworthy that despite the model's decision-making inclination toward a particular set of 8 frames, the maximum difference in weights does not exceed 0.12 when assigning temporal attention weights to the 32-frame continuous image sequences of the 5 behaviors. Thus, the model did not entirely neglect the other 24 frames, which still provided additional contextual information for formulating a holistic understanding of the behaviors. Consequently, the 3D-ResNet50-TSAM model demonstrated heterogeneity and bias in its attention allocation when recognizing different behaviors; nonetheless, the entire 32-frame image sequence contributed to the model's inferential judgment in a relatively equitable manner. These findings are pivotal for elucidating the decision-making process of deep learning models in broiler behavioral recognition tasks.

Building on the aforementioned analysis, this study further employs the Gradient-weighted Class Activation Mapping (Grad-CAM) technique to visualize the spatial attention of the 8 frames with the highest temporal attention weights for each behavior, as shown in Figure 5. Within each subfigure, the first row displays the original image frames, while the second row shows the corresponding heatmaps. The 3D-ResNet50-TSAM model focused more on the warmer color regions (red and yellow) relative to cooler ones (blue and green), and the more a region tended towards the color red, the greater its influence on the model's decision-making. For the activity behavior, the model primarily focused on the regions of the broiler's head and feet; for the spreading wing behavior, the model mainly focused on the skin areas exposed after the wings were spread; for the resting behavior, the model showed attention to the overall region of the broiler; for the feeding behavior, the model significantly focused on the broiler's head and the adjacent region of the feeder; and for the drinking behavior, the model similarly emphasized the broiler's head and the part of the drinker it contacted. These focus regions identified by the model showed high consistency with the regions humans paid attention to when recognizing these behaviors, further validating the 3D-ResNet50-TSAM model's reasonableness in behavior cognition.

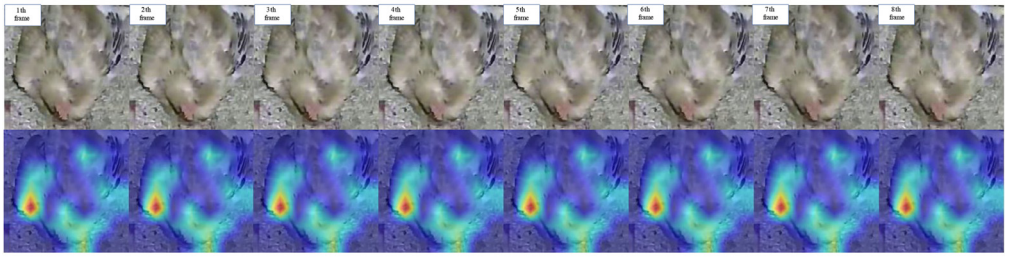
To further illustrate the effectiveness of the proposed 3D-ResNet50-TSAM, it was compared with commonly used 3D convolutional networks in the field of video behavior recognition, including C3D, TSN, TSM, TSM-



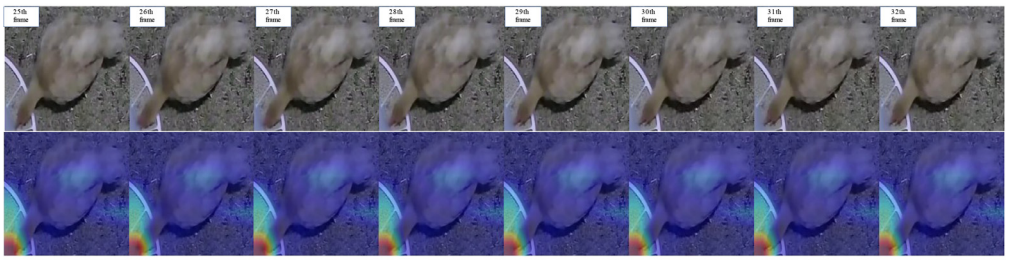
(a) Activity



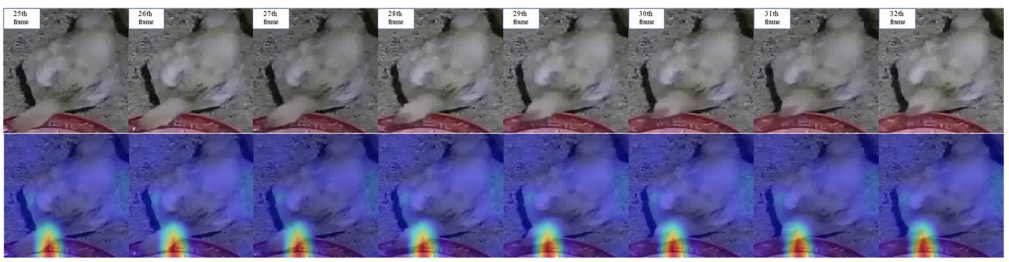
(b) Spreading wing



(c) Resting



(d) Feeding



(e) Drinking

Figure 5. Spatial attention visualization results of 5 broiler behaviors.

NL, I3D, and I3D-NL, with results presented in Table 7 (Tran et al., 2015; Wang et al., 2016; Carreira and Zisserman, 2017; Lin et al., 2019). The C3D model had *Accuracy*, *Precision*, *Recall*, and *F*1 *score* of 90.84, 90.74, 91.08, and 90.91%, respectively, which were lower than the other models, but it had the advantage in inference speed, at only 15 ms. The 3D-ResNet50-TSAM model performed the best among all the compared models, reaching the highest in *Accuracy*, *Precision*, *Recall*, and *F*1 *score* values at 97.84, 97.72, 97.65, and 97.68%, respectively. The TSM-NL model ranked second in performance, with an *Accuracy* only

Table 7. Performance comparison of different behavior recognition models.

Model	Backbone	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*F1 score* (%)	Inference time (ms)
C3D	-	90.84	90.74	91.08	90.91	15
TSN	ResNet-50	93.11	93.01	93.20	93.10	516
TSM	ResNet-50	96.29	96.11	96.47	96.29	620
TSM-NL	ResNet-50	96.87	96.67	97.05	96.86	651
I3D	3D ResNet-50	92.25	92.08	92.47	92.27	64
I3D-NL	3D ResNet-50	94.48	94.31	94.70	94.42	77
Ours	3D ResNet-50-TSAM	97.84	97.72	97.65	97.68	146

0.97% lower than the 3D-ResNet50-TSAM model, but with an inference time approximately 4.5 times longer. The inference times of the I3D and I3D-NL models were close to each other; however, their *Accuracy*, *Precision*, and other metrics were somewhat lacking compared to the 3D-ResNet50-TSAM model. Overall, although the 3D-ResNet50-TSAM model was not as fast in inference speed as the C3D, I3D, and I3D-NL models, it exhibited a clear advantage in comprehensive performance and could provide stable and reliable results for the broiler behavior recognition.

Analysis of the BBRS's Overall Performance

Effects of Tracking ID Switches on BBRS Accuracy

In the BBRS, the final accuracy of behavior recognition results was influenced by multiple factors, including the performance of the improved YOLOv8s detector, the Bytetrack tracker, and the 3D-ResNet50-TSAM behavior recognition model, all of which had been discussed in detail in the aforementioned sections. This section addressed the impact on behavior recognition results due to the issue of tracking ID switches, which could disrupt the 3D-ResNet50-TSAM model from receiving a complete 32-frame image sequence.

According to the behavior recognition logic of the BBRS, the tracker provides the behavior recognition model with a continuous 32-frame image sequence (1.28 s). If the target broiler's ID switches within this time window, the behavior recognition model will be forced to infer based on an incomplete image sequence prior to the switch. In light of this issue, this study evaluated the performance of the 3D-ResNet50-TSAM model in handling incomplete image sequences, with results shown in Figure 6. The longer the duration of the continuous image sequence of the target broiler received by the 3D-ResNet50-TSAM model, the higher the corresponding accuracy. When the model received approximately 80% of the complete image sequence (around 1.02 s or 26 continuous frames), the *Accuracy* of the 3D-ResNet50-TSAM model could reach 93.98%. The model's accuracy when handling a 1.15 s sequence was almost equivalent to that of a full sequence, but dropped significantly when only a 0.5 s sequence was available. Thus, for practical application of the BBRS, ensuring that its tracker consistently tracked the target broiler within the time window for at least 1.02 s could help to maintain a high accuracy level in behavior recognition.

Effects of Various Broiler Ages on BBRS Accuracy

The BBRS consists of 3 components: the improved YOLOv8s detector, the Bytetrack tracker, and the 3D-ResNet50-TSAM model. Among these components, the improved YOLOv8s detector and the 3D-ResNet50-TSAM model are deep learning models that require broiler images and videos for training. Therefore, during the development of BBRS, the object recognition dataset and the behavior recognition dataset were used to train and test the deep learning models within the system. These datasets were created using images of broilers ranging from 20 to 33-days-old throughout the experimental period. The inclusion of data from broilers at various ages during the training phase ensured that the BBRS could generalize well to different growth stages within the 20 to 33-day window. To further clarify the effect of various broiler ages on BBRS accuracy, the accuracy of the BBRS at 14 different broiler ages was evaluated (Figure S5). The result showed that the accuracy curve remained relatively stable, with minor fluctuations around the overall accuracy of 97.84%. This consistency reflected the robustness of the BBRS in recognizing broiler behaviors across different ages.

Effects of Various Broiler Speeds on BBRS Accuracy

The performance of BBRS was evaluated under 2 scenarios: slow movement and fast movement of broilers. Assuming linear movement of broilers within a unit of time, the average speed of each broiler was calculated for two 10-s videos (Video 1 and Video 2) using the tracking results from the Bytetrack tracker, where all

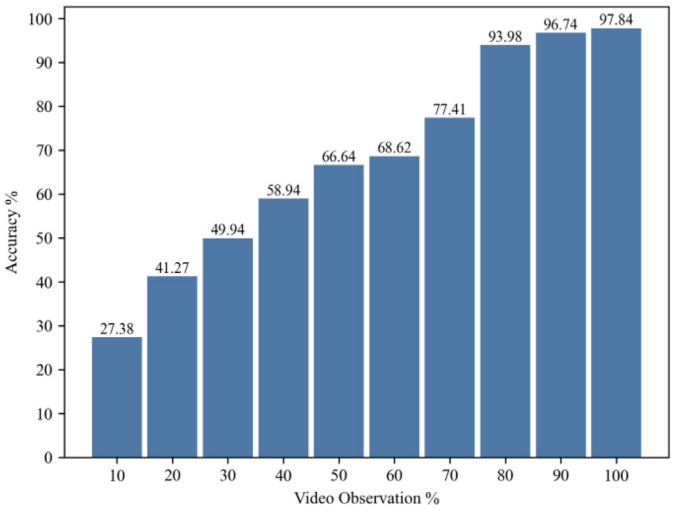


Figure 6. Behavior recognition results using incomplete image sequences.

Table 8. The average speed of each broiler tested in 2 videos.

	ID	1	2	3	4	5	6	7	8	9	10
Average speed (mm/s)	Video 1	27.5	26.7	4.2	16.9	0.9	5.6	13.0	10.3	16.3	1.5
	Video 2	95.0	81.0	26.4	4.5	44.8	9.5	14.4	8.1	75.6	19.8

broilers moved at a slow and fast pace, respectively. The results of the average speed for each broiler in both videos are shown in Table 8. All broilers moved slowly in Video 1, whose average speed did not exceed 27.5 mm/s, and BBRS provided satisfactory behavior recognition results within each time window. In contrast, the broilers with IDs 1, 2, and 9 in Video 2 exhibited rapid movements, whose average and maximum speeds reached 95.0 and 987.3 mm/s, 81.0 and 632.3 mm/s, and 75.6 and 1,045.5 mm/s, respectively. BBRS also succeeded in tracking these rapidly moving broilers and correctly identified their behavior as 'activity'. The results demonstrated that BBRS was capable of recognizing the behaviors of both slow-moving and fast-moving broilers. A demo video showcasing the recognition results of BBRS in both "Move slowly" and "Move fast" scenarios can be found at: https://www.bilibili.com/video/BV1Z7421K7sA/ or https://youtu.be/6W2z9XBvJHk.

Effects of Broiler Behavior Changes Within 32 Frames on BBRS Accuracy The decision to use a 32-frame time window for behavior recognition in the BBRS was based on experimental observations. The 5 broiler behaviors—activity, spreading wing, resting, feeding, and drinking—can be effectively expressed within this time window. During the training of the behavior recognition component, the 3D-ResNet50-

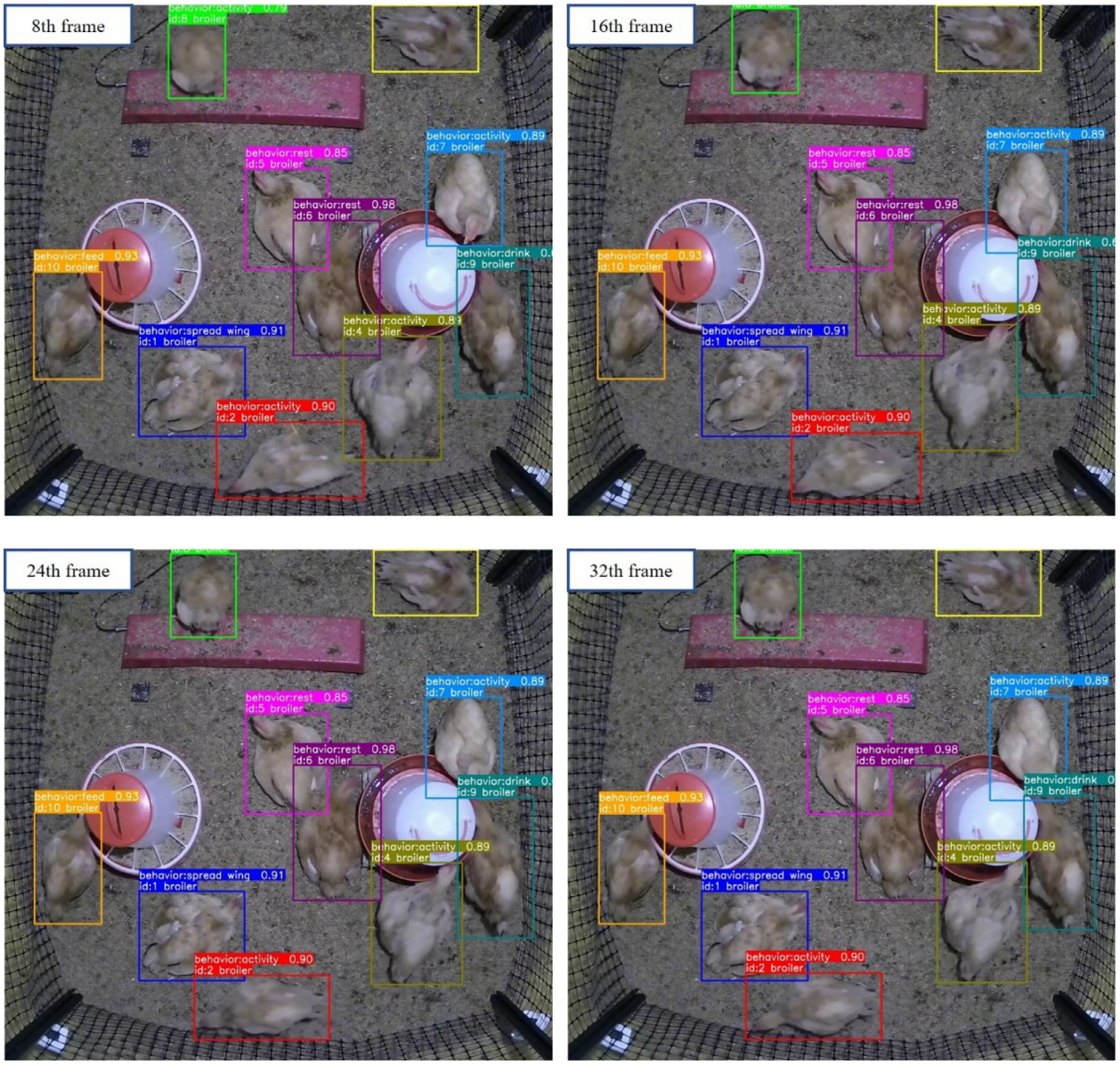


Figure 7. Visualization of the BBRS inference results within a time window.

TSAM model, all video segments used featured consistent behavior within the 32-frame window. However, maintaining consistent behavior during this period was not a strict requirement for the BBRS to function properly. The 3D-ResNet50-TSAM component was designed to learn spatiotemporal features from continuous image sequences, meaning that even if a broiler behavior changed within the 32-frame window, the model can still capture this change through its spatiotemporal attention module. According to the design logic of BBRS, the system processes each broiler video sequence in 32-frame increments, and even if a behavior change occurs within these frames, the system outputs a single behavior recognition result for the entire sequence, rather than separats results for before and after the behavior change.

Considering that 32 frames correspond to 1.28 seconds, it is possible for a broiler's behavior to change once within this brief period, such as transitioning from feeding to activity. A test was conducted to determine how BBRS responds when behaviors change within this time window and each behavior has sufficient time for complete expression before and after the change. The results indicated that if a broiler behavior transitioned from resting to activity or spreading wing, the BBRS output the initial behavior (resting). Similarly, if a broiler behavior transitioned from activity to resting, spreading wing, drinking, or feeding, the system output the initial behavior (activity). For transitions from spreading wing to resting or activity, the initial behavior (spreading wing) was recognized. However, if a broiler behavior transitioned from feeding or drinking to activity, the system output the final behavior (activity). These results demonstrated the influence of the TSAM in the model's decision-making process. Specifically, for the behaviors activity, spreading wing, and resting, the model tended to focus more on the first 8 frames of the image sequence, while for drinking and feeding, it concentrated more on the last 8 frames. It is crucial to emphasize that the likelihood of behavior changes occurring within such a short time window is inherently low, considering the nature of broiler behaviors and their actual expression patterns. Therefore, this does not undermine the effectiveness of BBRS in performing end-to-end spatiotemporal behavior recognition tasks.

Figure 7 presents the behavior recognition results of the BBRS for a continuous 32-frame image sequence. For clarity of demonstration, frames numbered 8, 16, 24, and 32 from the sequence had been specifically illustrated. The behaviors of broilers in this sequence included activity, spreading wing, resting, feeding, and drinking, allowing for a clear visual representation of different behaviors expressed by broilers with corresponding IDs.

To effectively deploy the BBRS on mini computers with limited computational power (e.g., edge devices) in the future, 2 avenues of optimization have been proposed in this research. Firstly, enhancing the tracker's long-range tracking capability in complex scenes to reduce the number of ID switches during tracking, and optimizing the behavior recognition model architecture to develop models capable of processing different effective video durations. Secondly, the conception of a lightweight model should help improve its overall efficiency, reduce the computational power demands, and result in lower deployment costs.

CONCLUSIONS

The study proposed an end-to-end method for automatically and accurately recognizing multiple simultaneous behavioral events of cage-free broilers in videos by Broiler Behavior Recognition System (**BBRS**) based on spatiotemporal feature learning. The BBRS consisted of 3 main components: the improved YOLOv8s detector, the Bytetrack tracker, and the 3D-ResNet50-TSAM model. The improved YOLOv8s detector exhibited an outstanding performance in multi-broiler detection tasks by integrating the MPDIoU to identify varying sizes and postures of broilers in the same frame of videos. The $F1$ *score*, *Precision*, *Recall*, and mAP@0.5 of the improved YOLOv8s detector reached 99.77%, 99.73%, 99.81%, and 99.50%, respectively. The Bytetrack tracker could stably track each identified broiler and acquire its image sequence of 32 continuous frames as input for the 3D-ResNet50-TSAM model. The mean *MOTA* and *IDF*1 *score* of the Bytetrack tracker were 93.89% and 87.55% at different occlusion levels. The 3D-ResNet50-TSAM model demonstrated high stability by integrating a temporal-spatial attention module, which was used to learn the spatiotemporal features from its image sequence and enhance inference ability in the case of its image sequence less than 32 continuous frames due to its tracker ID switching. The *Accuracy*, *Precision*, *Recall*, and $F1$ *score* of the 3D-ResNet50-TSAM model were 97.84, 97.72, 97.65, and 97.68%, respectively. The BBRS showed satisfactory inference ability with an *Accuracy* of 93.98% when the tracker ID switched and 26 continuous frames of the tracked broiler were received by the 3D-ResNet50-TSAM model. Furthermore, BBRS was capable of recognizing the behaviors of broilers across various ages and speeds. These results indicated that the BBRS was accurate and reliable in the task of end-to-end cage-free multi-broiler behavior recognition.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the National Key R&D Program of China (2023YFD2000801). Any opinions, findings or conclusions expressed in this publication are those of the authors and do not necessarily reflect the views of Zhejiang University. The trade and manufacturer names must be reported factually on the available data.

DISCLOSURES

The authors declare no conflicts of interest.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.psj.2024.104314.

REFERENCES

Carreira, J., and A. Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. Pages 4724-4733 in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

China Animal Agriculture Association. 2022. China Intelligent Animal Husbandry Development Report 2022. China Agriculture Press, Beijing.

Fang, C., T. M. Zhang, H. K. Zheng, J. D. Huang, and K. X. Cuan. 2021. Pose estimation and behavior classification of broiler chickens based on deep neural networks. Comput. Electron. Agric. 180:105863.

Handan-Nader, C., and D. E. Ho. 2019. Deep learning to map concentrated animal feeding operations. Nat. Sustain. 2:298–306.

Hara, K., H. Kataoka, Y. Satoh, and IEEE.. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? Pages 6546-6555 in 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

He, P. G., Z. H. Chen, H. W. Yu, K. Hayat, Y. F. He, J. M. Pan, and H. J. Lin. 2022. Research progress in the early warning of chicken diseases by monitoring clinical symptoms. Appl. Sci.-Basel 12:5601.

He, K. M., X. Y. Zhang, S. Q. Ren, J. Sun, and IEEE.. Deep residual learning for image recognition. Pages 770-778 in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ji, H., G. Teng, J. Yu, Y. Wen, H. Deng, and Y. Zhuang. 2023. Efficient aggressive behavior recognition of pigs based on temporal shift module. Animals 13:2078.

Kirillov, A., E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick. 2023. Segment anything. Pages 3992-4003 in 2023 IEEE/CVF International Conference on Computer Vision (ICCV).

Lin, J., C. Gan, and S. Han. 2019. TSM: temporal shift module for efficient video understanding. Pages 7082-7092 in 2019 IEEE/CVF International Conference on Computer Vision (ICCV).

Liu, Y. Y., X. Cao, B. B. Guo, H. J. Chen, Z. C. Dai, and C. W. Gong. 2023. Research on detection algorithm about the posture of meat goose in complex scene based on improved YOLO v5. J. Nanjing Agric. Univ. 46:606–614.

Ma, S. L., Xu, Y., 2023. MPDIoU: a loss for efficient and accurate bounding box regression.

Misra, D. 2020. Mish: a self regularized non-monotonic activation function. British Machine Vision Conference.

Nasiri, A., J. Yoder, Y. Zhao, S. Hawkins, M. Prado, and H. Gan. 2022. Pose estimation-based lameness recognition in broiler using cnn-lstm network. Comput. Electron. Agric. 197:106931.

Rakai, L., H. S. Song, S. J. Sun, W. T. Zhang, and Y. N. Yang. 2022. Data association in multiple object tracking: a survey of recent techniques. Expert Syst. Appl. 192:116300.

Subedi, S., R. Bist, X. Yang, and L. L. Chai. 2023. Tracking pecking behaviors and damages of cage-free laying hens with machine vision technologies. Comput. Electron. Agric. 204:107545.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. Pages 4489-4497 in 2015 IEEE International Conference on Computer Vision (ICCV).

Volkmann, N., C. Zelenka, A. M. Devaraju, J. Bruenger, J. Stracke, B. Spindler, N. Kemper, and R. Koch. 2022. Keypoint detection for injury identification during turkey husbandry using neural networks. Sensors 22:5188.

Wang, L. M., Xiong, Y. J., Wang, Z., Qiao, Y., Lin, D. H., Tang, X. O., Van Gool, L., 2016. Temporal segment networks: towards good practices for deep action recognition. Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.). Springer International Publishing, Cham, pp. 20-36.

Wang, X. L., R. Girshick, A. Gupta, and K. M. He. 2018. Non-local neural networks. Pages 7794-7803 in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Wang, J., N. Wang, H. H. Li, and Z. H. Ren. 2020. Real-time behavior detection and judgment of egg breeders based on YOLO v3. Neural Comput. Appl. 32:5471–5481.

Xiao, L. F., K. Y. Ding, Y. W. Gao, and X. Q. Rao. 2019. Behavior-induced health condition monitoring of caged chickens using binocular vision. Comput. Electron. Agric. 156:254–262.

Xiao, D. Q., R. L. Zeng, M. Zhou, Y. G. Huang, and W. C. Wang. 2023. Monitoring the vital behavior of magang geese raised in flocks based on dh-yolox. Transact. Chin. Soc. Agric. Eng. 39:142–149.

Yang, X., R. Bist, S. Subedi, Z. Wu, T. Liu, and L. Chai. 2023a. An automatic classifier for monitoring applied behaviors of cage-free laying hens with deep learning. Eng. Appl. Artif. Intell. 123:106377.

Yang, X., R. Bist, S. Subedi, and L. L. Chai. 2023b. A deep learning method for monitoring spatial distribution of cage-free hens. Art. Intellig. Agric. 8:20–29.

Zhang, Y. F., P. Z. Sun, Y. Jiang, D. D. Yu, F. C. Weng, Z. H. Yuan, P. Luo, W. Y. Liu, and X. G. Wang. 2022. Bytetrack: Multi-Object Tracking by Associating Every Detection Box. Pages 1−21 in ECCV 2022. S. Avidan, G. Brostow, M. Cisse, G. M. Farinella and T. Hassner, eds. Springer International Publishing PT XXII.

Zhao, C. Z. 2019. State-of-the-art and recommended developmental strategic objectivs of smart agriculture. Smart Agric. 1:1–7.

Zheng, Z. H., P. Wang, W. Liu, J. Z. Li, R. G. Ye, and D. W. Ren. 2020. Distance-iou loss: faster and better learning for bounding box regression. Pages 12993-13000 in 34th AAAI Conference on Artificial Intelligence /32nd Innovative Applications of Artificial Intelligence Conference / 10th AAAI Symposium on Educational Advances in Artificial Intelligence.