Computers and Electronics in Agriculture 224 (2024) 109243

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Deep neural network with adaptive dual-modality fusion for temporal aggressive behavior detection of group-housed pigs

Kai Yan [a], Baisheng Dai [a,b,*], Honggui Liu [c], Yanling Yin [a], Xiao Li [d], Renbiao Wu [e], Weizheng Shen [a,**]

[a] College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China
[b] Key Laboratory of Pig-breeding Facilities Engineering, Ministry of Agriculture and Rural Affairs, Harbin 150030, China
[c] College of Animal Science and Technology, Northeast Agricultural University, Harbin 150030, China
[d] China Agriculture Press, Beijing 100125, China
[e] Sichuan Xinmu Hui Technology Co., Ltd., Chengdu 610000, China

ARTICLE INFO

Keywords:
Aggressive behavior
Group-housed pigs
Temporal behavior detection
Dual-modality fusion

ABSTRACT

The frequent occurrence of pig aggressive behaviors in intensive group-housed environment seriously affects pig health, welfare and farms economy. Accurate detection of the occurring and temporal interval of aggressive behaviors is important for pig farming. The study aimed to develop an automatic temporal aggressive behavior detection method based on deep neural network. This network mainly consists of three modules, i.e., aggression feature extraction, adaptive dual-modality fusion and aggression temporal proposal generation. First, RGB data and optical flow data was used to extract the spatial and motion information of pig aggressive behaviors. Second, a modality attention and a temporal attention were specifically designed to adaptively fuse features of different modalities. Third, an anchor-free aggression temporal proposal generation strategy was applied to generate aggression proposals, which indicate the start and end times of aggressive behavior. To evaluate the proposed method, a behavior dataset containing 216 videos and 642 annotations was constructed. On the test set, this method achieves an AP value of 68.0 %, an AR value of 77.8 % in average number of proposals at 100. To test this method in practical application, our method was conducted on an additional 90 min untrimmed surveillance video and effectively predicted the real aggression instances. The results demonstrated that it can meet the practical needs of intelligent monitoring in pig farming and analysis in animal behavior research. We shared our temporal aggressive behavior detection dataset at https://github.com/IPCLab-NEAU/Temporal-Aggressive-Behavior-Detection for precision livestock farming research community.

1. Introduction

In intensive pig farms, mixing unfamiliar pigs is a standard procedure during the wean-to-market period (Gonyou 2001). After mixing, pigs will engage in frequent aggressive behavior to establish a new social hierarchy (de Groot et al., 2001, Coutellier et al 2007). Negative consequences of aggressive behavior include skin lesions, lameness and social stress, which can affect the health, welfare of pigs and farms economy (McGlone 1985). Therefore, real-time monitoring and timely discovering aggressive behaviors could effectively reduce wound infections and even deaths in pigs, which is of great significance for achieving healthy pig farming and improving farms economy (He et al 2016).

In the past, pig farms mainly relied on human observation to monitor aggressive behavior among pigs which was subjective and ineffective. Considering the objectivity and accuracy of computer vision, a number of researchers have explored vision-based methods for automatically recognizing aggressive behavior from surveillance video. In early stage, the recognition methods usually used manually defined features to characterize aggressive behavior. These features included geometrical features (e.g., distance between pigs, connected area and adhesion index) and motion features (e.g., mean motion intensity, activity index, speed, acceleration, kinetic energy and motion pixel) (Viazzi et al. 2014, Oczak et al. 2014, Lee et al. 2016, Chen et al. 2017, Chen et al. 2018,

* Corresponding author at: College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China.
** Corresponding author.
E-mail addresses: bsdai@neau.edu.cn (B. Dai), wzshen@neau.edu.cn (W. Shen).

https://doi.org/10.1016/j.compag.2024.109243
Received 13 July 2023; Received in revised form 7 July 2024; Accepted 10 July 2024
Available online 14 July 2024

Table 1
Definition of aggressive behavior (reference from our previous work (Gao et al. 2023)).

Behavior classification		Behavior description
Biting	Bite tail	The aggravated pig bites the victim's tail
	Bite ear	The aggravated pig bites the victim's ear
	Bite body	The aggravated pig bites the victim's body (except the tail and ears)
Hitting	Head to head hit	The aggravated pig hits the victim's head with its head
	Head to body hit	The aggravated pig hits the victim's body with its head
Trampling	The aggravated pig tramples on the victimized pig with its feet	
Chasing	Chase between parties after aggressive behavior	

Chen et al. 2019). With the development of deep learning, extracting spatio-temporal features of aggressive behavior through deep neural network became mainstream. Gao et al. (2019) proposed an integrated multi-scale feature fusion 3D convolutional neural network (CNN) to learn spatial and temporal features, and developed an end-to-end model for pig aggression recognition. Gan et al., (2021a) used a key point detector to detect piglet key points and obtained pig movement information through a tracking unit which was used for behavioral classification. Chen et al. (2020) and Gao et al. (2023) used a deep neural network framework that successively extracts spatial and temporal features to recognize aggressive behavior, in which convolutional neural network was used to extract spatial features, and recurrent neural network (e.g., LSTM and GRU) was used to aggregate temporal features.

Most existing studies on aggressive behavior recognition focuses on behavior classification using short video clips, which assumes implicitly that aggressive behavior was manually trimmed from surveillance video during both training and testing (Gao et al. 2019, Chen et al. 2020, Gan et al. 2021a, Gao et al. 2023). Despite the great improvements in behavior classification, problem still persists. The trimming strategy simplifies the real-time video behavior analysis but is difficult to apply in real farming environment, considering that surveillance video is untrimmed which contains multiple aggression instances as well as a large number of background behaviors. To address this problem, a temporal behavior detection method was applied to monitor aggressive behavior. This method was able to predict the temporal interval of each behavior in untrimmed video and categorize each behavior, which has potentially research and application value than previous studies for intelligent monitoring in pig farming (Perez et al. 2023). Current temporal behavior detection mainly focused on learning actionness of each frame (Lin et al. 2019, Yu et al. 2021, Su et al. 2021), adjustment of pre-defined anchors (Xu et al. 2017, Liu et al. 2020, Yoon et al, 2020, Xu et al, 2020) or regression distance with anchor-free strategy (Lin et al. 2021, Zhang et al. 2022, Shi et al. 2023). Actionness strategy processes video feature sequences in a bottom-up approach. It predicts action and boundary probabilities for each frame. Based on the prediction scores, behavior proposals are generated. This strategy requires an external classifier for behavior classification. Pre-defined anchors strategy presets multi-scale anchors to obtain candidate proposals, and performs behavior classification and regression based on these proposals. Due to variations in the duration of aggression instances, pre-defined anchor strategy requires a large amount of computation when generating dense candidate proposals. Anchor-free strategy utilizes temporal pyramid block to generate multi-scale features which used to directly regress the boundaries for each frame. This strategy is more flexible compared to others.

Based on the above motivations, we explored the application of temporal aggressive behavior detection framework with anchor-free strategy that locates time interval of aggression by extracting spatial and motion information from surveillance video. Spatial information is generally extracted from RGB data directly by deep neural network, while motion information was commonly modeled with motion history images (Viazzi et al. 2014), frame difference method (Chen et al. 2019) and optical flow (Yang et al. 2019, Yang et al. 2020, Gan et al. 2021b). Specially, since optical flow is less affected by background noise, more and more researchers obtained motion information from optical flow data. Spatial and motion information were usually combined to capture behavior pattern. Current research fused spatial and motion information by simply concatenating or weighting them (Yang et al. 2020, Gan et al. 2021b, Bati and Ser, 2023). However, the spatial and motion information were entangled with each other. Roles of different modality were not sufficiently considered by these studies.

In this work, we proposed a temporal aggressive behavior detection method to monitor group-housed pigs' behaviors based on deep neural network, and specially designed an adaptive dual-modality fusion module to fuse spatial and motion information for extracting aggressive pattern. Particularly, an aggression feature extraction module was first applied to obtain spatial and motion information. The adaptive dual-modality fusion module then fused them by integrating a modality attention and a temporal attention. Modality attention was designed to extract modality weight information within the current segment. Temporal attention was proposed to enhance the features of key frames. Finally, an aggression temporal proposal generation module was used to generate proposals, including confidence, start frame and end frame of aggressive behavior instance.

The remaining parts of this paper are presented as follows: In section 2, the dataset and statistical information of video data are introduced; In Section 3, the overview of the temporal aggressive behavior detection method is showed; The detection results of aggressive behavior are discussed in Section 4 and the conclusion is provided in Section 5.

2. Materials

2.1. Animals and experimental installations

The video data was collected from Harbin HongFu Pig Farm. The pig pen size was 4.3 m × 2.3 m, the rearing density is about 10 pigs per pen and there were feeding troughs and nipple drinkers inside the pen. The pig breed was Large white × Landrace piglets. The age of the pigs is 35–42 days. Due to the frequent aggressive behavior within three days after pigs were mixed group (Spoolder et al., 2000), the video data within 72 h after mixing were recorded.

The Hikvision DS-2CD3345D-I camera was used to record video with vertical view. The camera was placed above the pen at the height of 2.3 m relative to the ground. The resolution of the camera was 2560 × 1440 pixels, the frame rate was 25fps, and the video data was stored in.MP4 format. In subsequent processing, the frame rate was downsampled to 10 fps with interframe sampling, and the image in the video was adjusted to 112 × 112 pixels (Lin et al 2021) using bilinear interpolation to save processing power and memory in subsequent experiments.

2.2. Definitions of aggressive behaviors

Aggressive behavior in pigs involves the interaction of several pig and is a complex and progressive behavior. At the beginning of aggressive interactions, pigs typically make preliminary bites by sniffing and pawing, followed by more forceful bites and hits that typically occur as aggressive behavior increases. More serious aggressive biting which include ear biting, tail biting, and body biting will occur in the most intense stages. In addition, chasing and trampling are also common aggressive behaviors in pigs (Kongsted, 2004, Turner et al., 2006). The aggressive behaviors of group-housed pigs defined in this dataset mainly include biting, hitting, trampling and chasing. Biting can be subdivided into tail biting, ear biting and body biting, and hitting is subdivided into head hitting and body hitting. The specific descriptions are shown in Table 1.

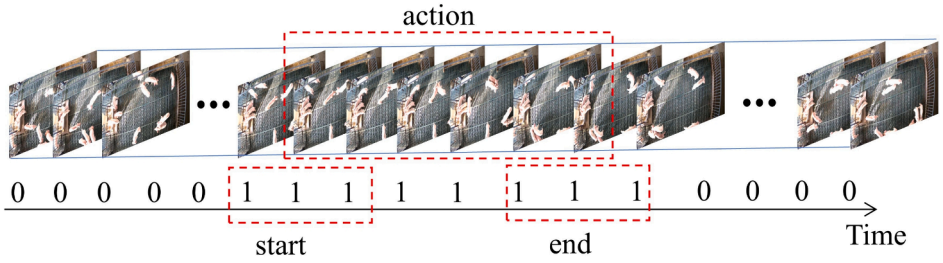

Fig. 1. Specific label setting for the dataset. Utilize annotation information to determine start and end labels by selecting adjacent frames.

Table 2
Annotated file information.

Annotated file name	In-file information
video_info.csv	video name、video fps、sample fps、video count、sample count
annotation_info.csv	video name、behavior type、start time、end time、start frame、end frame

Table 3
Allocation of train and test set.

Dataset category	Number of videos	Number of instances
Train dataset	175	482
Test dataset	41	160

2.3. Data labelling and definitions of aggressive characteristics

Raw data is the surveillance video of pig farm within 72 h, each segment length is 90 min. To reduce the effect of a large number of background frames on aggressive pattern extraction, videos are trimmed into clips of variable size, from 1 to 10 min. Through manual supervision, the clips are ensured to contain at least one complete instance of live pig aggressive behavior. In addition, several raw data are preserved for analysis and testing of untrimmed video. The train set contains 175 segments and the test set contains 41 segments.

The annotation of dataset includes start and end frames of aggressive behavior. Specific label setting is shown in Fig. 1. Elan software and Python scripts were used for annotation in the experiment. The annotation information includes the start and end times of aggressive behaviors. Through scripts, we can convert the annotation data into file information as shown in Table 2 below.

On average, each video contains 2.7 segments of aggression annotation, and the distribution table of the dataset is shown in Table 3.

To effectively analyze aggressive behavior, we describe the dataset with inherent characteristics such as length of video (Video Length), duration of aggressive behavior (Instances Length) and number of aggressive behaviors per video (Instances Number). The results are shown in Fig. 2.

(1) Video Length: The length (in seconds) of each video segment was calculated and five categories were created: Extra Small (XS: [0,60]), Small (S: (60,120]), Medium (M: (120,180]), Large (L: (180, 240]), and Extra Large (XL: (>240)). The shortest video length was 40 s and the longest was 637 s. Small and Medium accounted for the majority of the total video segments at 70.7 %, indicating that most videos were between 1 and 3 min in video length.
(2) Instances Length: Instance Length is the duration of the aggressive behavior. The duration (in seconds) of the aggressive behavior contained in each instance was calculated and five categories were created: Extra Small (XS: [0, 5]), Small (S: (5, 10)), Medium (M: (10, 20)), Large (L: (20, 30)), and Extra Large (XL: (>30)). Among them, the shortest instance of aggressive behavior was 1 s and the longest was 492 s. The percentage of aggressive behavior up to 20 s in length was 89.2 %, and the percentage of aggressive behavior within 5 s and between 5 and 10 s was 36.6 % and 32.2 %, respectively.
(3) Instances Number: Instance Number is the number of aggressive behaviors per video. The number of instances of aggressive behavior (in units) per video was calculated and five categories were created: Extra Small (XS: [0,1]), Small (S: [2,4]), Medium (M: [5,8]), and Large (L: (>8)). Among them, each video contains at most 15 instances of aggressive behavior, and more than half of the videos contain 2 to 4 instances of aggressive behavior.

2.4. Optical flow data

Optical flow is a widely used approximation for estimating pixel motion between two consecutive frames based on apparent velocities of image motion (Dawkins et al., 2009). Many researchers have demonstrated the feasibility of extracting motion information from optical flow data (Gronskyte et al. 2016, Yang et al. 2019, Yang et al. 2020). The TV-

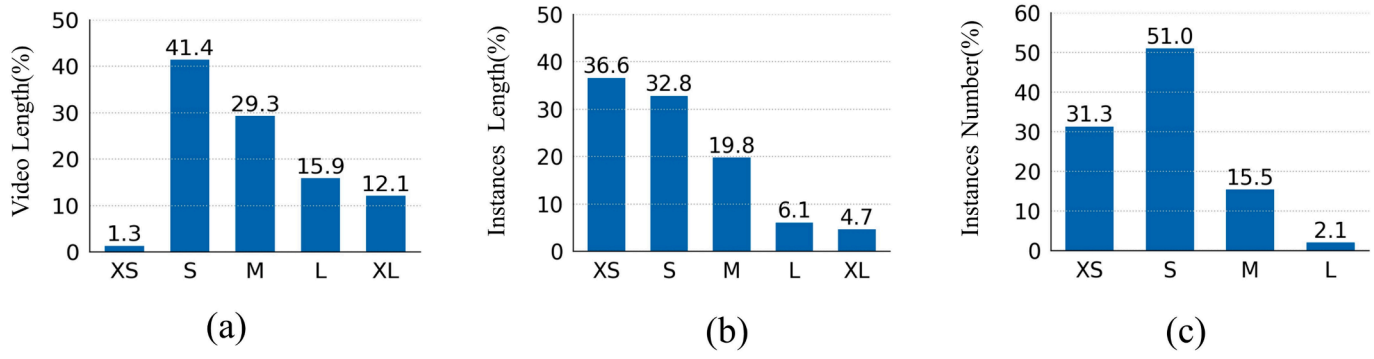

Fig. 2. Statistics of inherent characteristic within the dataset. (a) represent the video length. (b) represent the instance length of aggressive behavior. (c) represent the instance number per video.

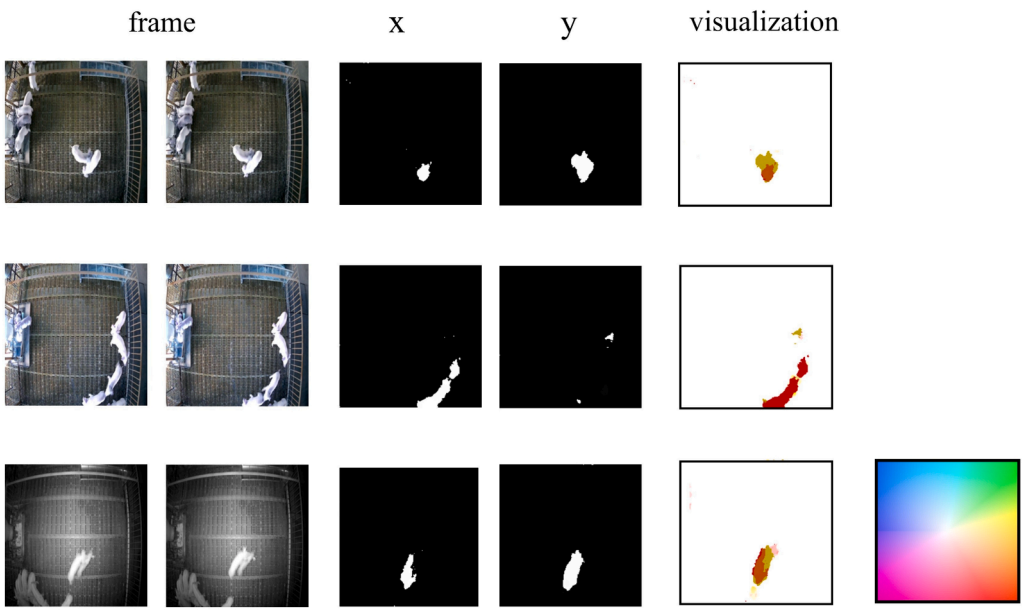

Fig. 3. Optical flow data between adjacent frames from surveillance video.

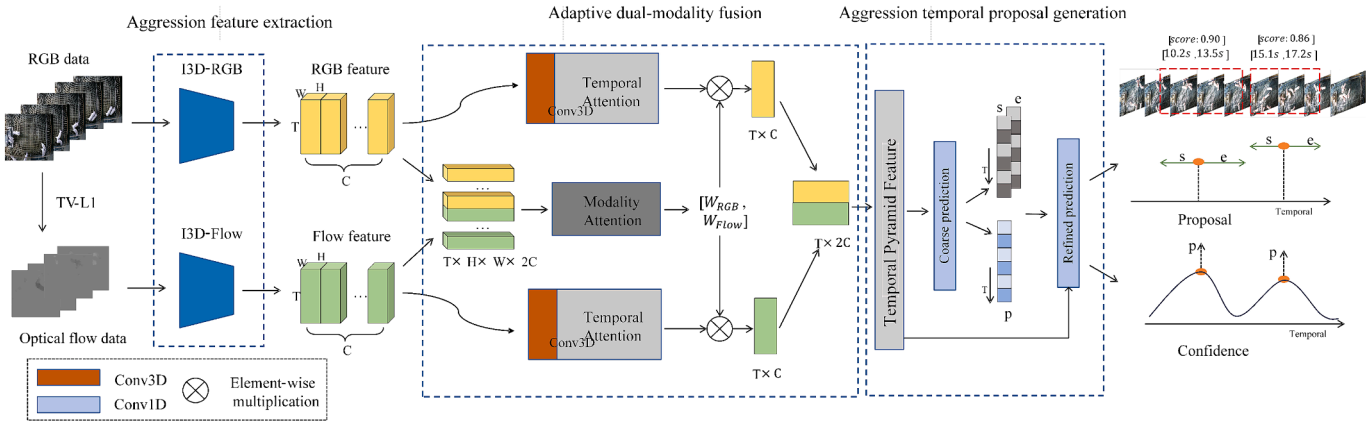

Fig. 4. Flowchart of the temporal aggressive behavior detection method.

L1 algorithm is used to obtain horizontal (x) and vertical (y) flow values. The flow values are constrained in the range [-20, 20] to remove noise, and were then normalized to the range 0 to 255 by a linear transformation. Two adjacent frames and their corresponding x and y optical flow trajectories are shown in Fig. 3. The grayscale values of pixels in the third and fourth columns represent the velocity vectors of each pixel. For easier observation, the optical flow images are overlaid and color-coded, with hue representing the direction of motion and saturation representing the intensity of motion.

3. Methods

3.1. Overview of temporal aggressive behavior detection

Aggressive behavior is an interactive activity that occurs between two or more pigs over a period of time. Different from previous studies, temporal aggressive behavior detection aims to predict temporal interval of each behavior in long or untrimmed videos and categorize each behavior. To detect aggressive behavior, it is necessary to extract spatial and motion features from video. Spatial features characterize the key regions and appearance of aggressive behavior (e.g., the relative position of the head and ears/tails during ear biting and tail biting). Motion features represent the movement information of pigs during behavior. We propose an adaptive dual-modality fusion module to fuse spatial and motion features. The overall method is shown in Fig. 4. The method consists of three steps. First, pre-trained I3D model (Carreira and Zisserman, 2017) was used to obtain aggressive behavioral features from RGB and optical flow data. Second, adaptive dual-modality fusion was proposed to fuse different modality features. In the end, aggression temporal proposal generation module was used to detect the occurring of aggressive behavior and locate the start and end frames, as well as the corresponding confidence scores.

3.2. Aggression feature extraction

In the dataset, aggressive behavior occurs alongside some background behaviors. Therefore, the aggression feature extraction module must accurately capture features among the aggressive behavior in video. For RGB and optical flow modalities, the I3D model which has trained on large RGB and optical flow datasets is used to extract aggression features. I3D model uses 3D convolution to extract features. 3D convolution can capture short-term spatio-temporal information between adjacent frames by simultaneously convolving both spatial and temporal dimensions. The model structure of I3D is illustrated in Fig. 5.

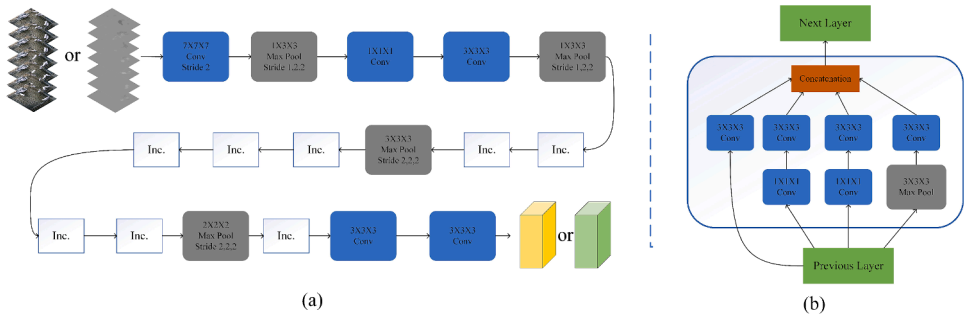

Fig. 5. The structure of aggression feature extraction module. (a) is the specific structure of I3D model. (b) is the internal structure of Inc. block.

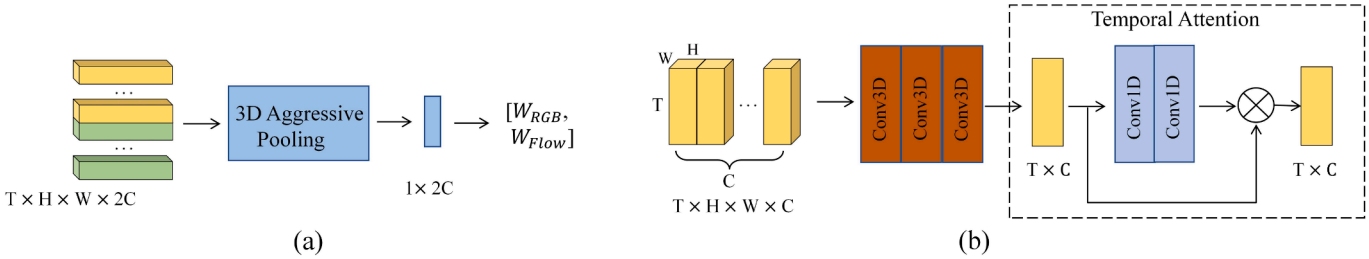

Fig. 6. Modality and temporal attention block structures. (a) is the modality attention block and (b) is the temporal attention block.

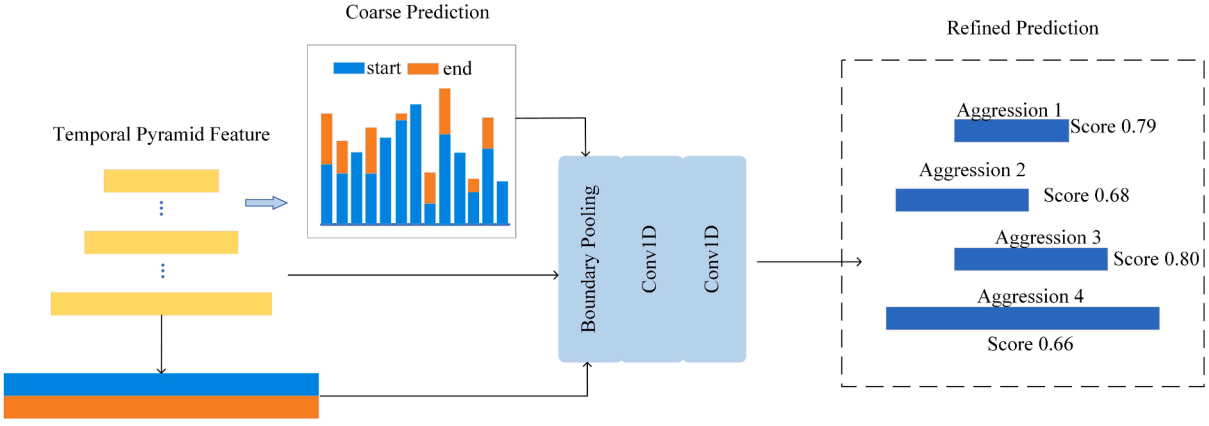

Fig. 7. The structure of aggression temporal proposal generation model AFSD.

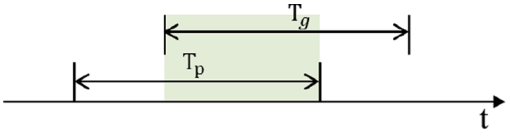

Fig. 8. Diagram of calculation method of tIoU.

Table 4
Experimental environment parameter.

Parameter settings	Version
Operating System	Windows 10
CPU	Intel(R) Core (TM) i7-7800X
RAM	DDR3 16G*12
GPU	NVIDIA GTX 1080Ti
Pytorch	1.6.0

Table 5
AP under different tIoU thresholds.

Method	AP(%)@tIoU					Avg. (%)
	0.1	0.2	0.3	0.4	0.5	
RGB	57.3	56.3	48.1	42.9	28.2	46.6
Flow	62.8	58.1	50.9	43.1	31.8	49.3
Fusion (concatenate)	67.3	61.7	54.0	44.7	35.0	52.5
Adaptive Fusion	68.0	65.8	58.8	49.3	35.7	55.5

Table 6
AR under different AN (Average Number).

Method	AR@5(%)	AR@10(%)	AR@100(%)	Num
RGB	60.5	70.8	75.8	1158
Flow	64.9	72.4	76.5	1571
Fusion (concatenate)	64.8	74.3	74.3	798
Adaptive Fusion	68.6	75.8	77.8	1065

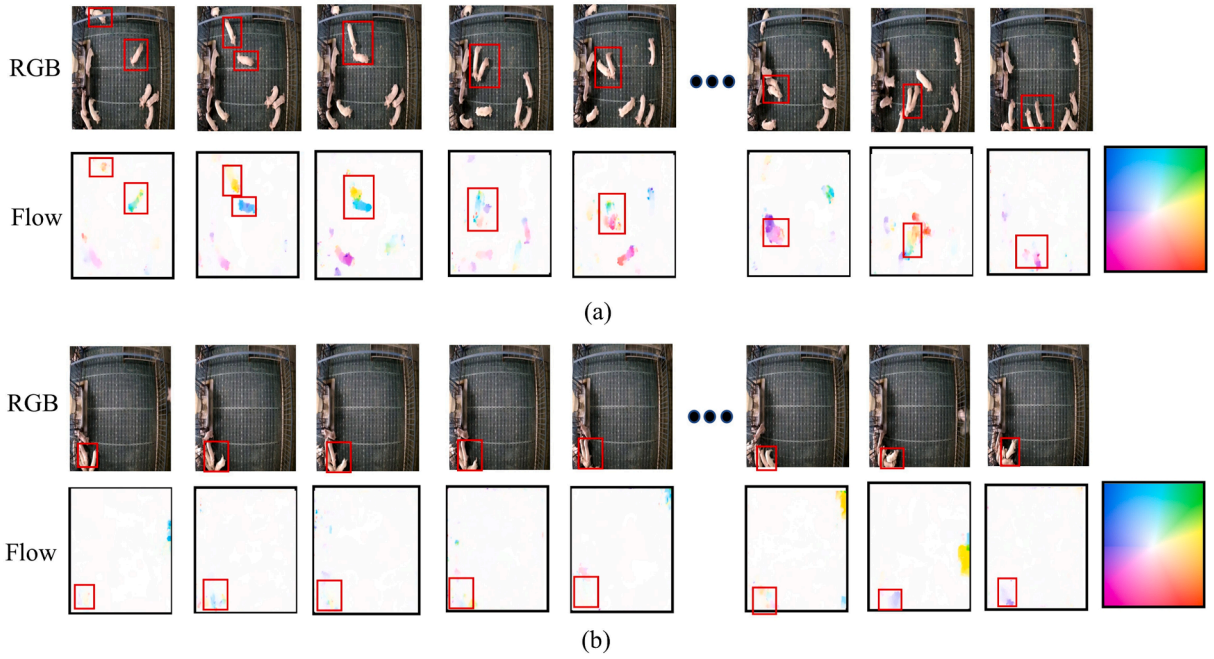

Fig. 9. Visualization results of RGB and flow modality data. (a) shows multiple motion behaviors occurring simultaneously in a large displacement aggressive behavior scene. (b) shows a segment of aggressive behavior which happened in dense pig herd.

The size of input video is 256 × 112 × 112 × 3 in RGB modality, while 256 × 112 × 112 × 2 in flow modality. Sliding window is used to extract feature with the stride 30 in training and 128 in testing. The size of RGB and optical flow modality features after aggression feature extraction module is T × H × W × C, which is specifically 64 × 7 × 7 × 832 in our experiments. T represents the time length after feature extraction. H, W represents the spatial height and width size of 7 × 7 after 3D convolution, and C represents the channel size of 832.

3.3. Adaptive dual-modality fusion

In previous processing, fusion was often performed by concatenating the different modality features directly or by assigning appropriate weights (e.g., RGB modal weight of 1 and optical flow modal weight of 1.5). However, for different behavior, the contributions provided by RGB and optical flow modalities are different. Recently, Wang et al. (2020) and Yang et al. (2022) has noticed the complementarity between the different modalities, and explored effective strategies to leverage different modality information in the field of video understanding. In order to adaptively fuse information from both modalities, an adaptive dual-modality fusion module is proposed.

The adaptive dual-modality fusion module consists of convolutional blocks, temporal attention block and modality attention block. Modality attention block extracts the modality weights within the current segment, and assigns weights to the different modality features. The temporal attention module focuses on the key frames where the aggressive behaviors occur. 3D convolutional block is used to compress the spatial features and aggregates local motion information. 1D convolutional block is utilized to merge the features from both modalities.

The schematic of modality attention block is shown in the Fig. 6(a). First, an aggression aware pooling layer is used to select representative aggressive behavior features for each frame position and suppresses background noise. Then, k most prominent temporal dimensions are selected as indices to extract the original features. The original features are averaged over k temporal dimensions, and a 3D convolution and softmax function are used to obtain the RGB modality weight and optical flow modality weight, which represent the current segment. The specific calculation process is described as follows.

$$f_t = 3DPooling\left(con\left(f_{rgb}, f_{flow}\right)\right) \tag{1}$$

$$F_k = \frac{1}{N_k}\sum_{t=1}^{N_k} max_{f_1,f_2,f_3\cdots f_k}(con(f_{rgb}, f_{flow}))\# \tag{2}$$

$$w_{rgb}, w_{flow} = softmax(conv3D(F_k))\# \tag{3}$$

where f_t represents the feature of current frame after pooling. f_{rgb} and f_{flow} represent the features of the RGB and Flow modalities respectively. N_k indicates the number of selected representative frames. w_{rgb} and w_{flow} are the final modality attention weights.

The temporal attention block focus on key frames in the sequence and reduces the interference from redundant frames. It can automatically locate, identify key frames and assign weight to each frame. The temporal attention module is shown in Fig. 6(b), and its purpose is to compress and aggregate the channel features of each frame through two 1D convolution layers.

Taking the RGB modality feature as an example, the specific calculation process is described as follows.

$$w = ReLU\left(conv1D\left(f_{rgb}\right)\right) \tag{4}$$

$$w_{att} = Sigmoid(conv1D(w))\# \tag{5}$$

$$f_{rgb} = w_{att}\hat{A}\cdot f_{rgb}\hat{A}\cdot w_{rgb}\# \tag{6}$$

where w_{att} is the weight of temporal attention, w_{rgb} is the weight of modality attention.

3.4. Aggression temporal proposal generation

Aggression temporal proposal generation module adopts AFSD (Anchor-Free Salient Detection) model (Lin et al 2021), which uses anchor-free localization strategy. Anchor-free strategy is more flexible for boundary localization and suitable for aggressive behavior with variable boundaries in group-housed pigs' surveillance video. For each frame in the temporal pyramid features, the strategy can directly regress

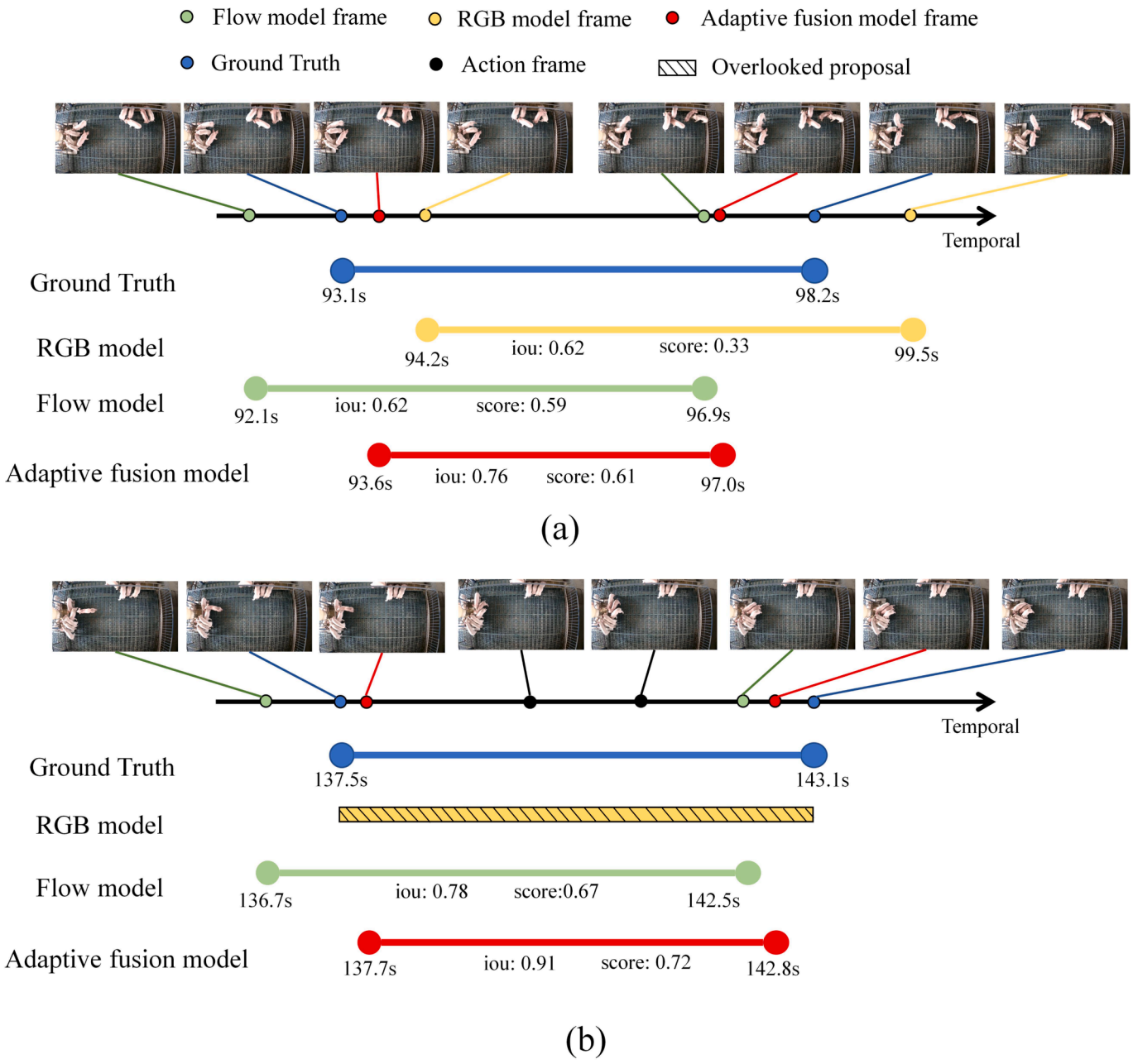

Fig. 10. Detection results on the test set. (a) and (b) show that the fusion results are more effective.

Table 7
Different attention block comparison experiment.

Model	Avg. AP (%)	AP@0.3 (%)	AP@0.5 (%)	AR@10 (%)
Fusion (concatenate)	52.5	54.0	35.0	74.3
Fusion (modality attention)	54.6	56.2	35.5	74.8
Fusion (temporal attention)	53.9	55.9	34.9	75.1
Fusion (both attention)	55.5	58.8	35.7	75.8

the temporal distance and classify behaviors. Compared to other strategies, it uses fewer proposals to achieve higher detection performance without external classifier. Aggression temporal proposal generation module uses the enhanced dual-modality fusion feature as input to obtain the start frame, end frame and the confidence of aggressive behavior, which includes temporal pyramid block and conv prediction blocks.

The AFSD model mainly consists of a multi-scale temporal pyramid block, a coarse prediction block, and a refined prediction block. The multi-scale temporal pyramid block generates multi-scale temporal features. The coarse prediction block initially predicts aggressive behavior proposals through a classification head and a regression head. The refined prediction block further refines aggressive behavior proposals by combining frame-level boundary features, coarse proposals and pyramid features at different scales. The structure of AFSD model is shown in Fig. 7.

The fused features encompass both spatial apparent and motion information from entire video, with a size of T×C. T represents the temporal length, and C denotes the channel dimension. Single frame contains the spatial position and apparent information of the pig, and the motion information of the pig is contained between multiple frames. Specifically, a temporal pyramid network with stacked 1D convolutions

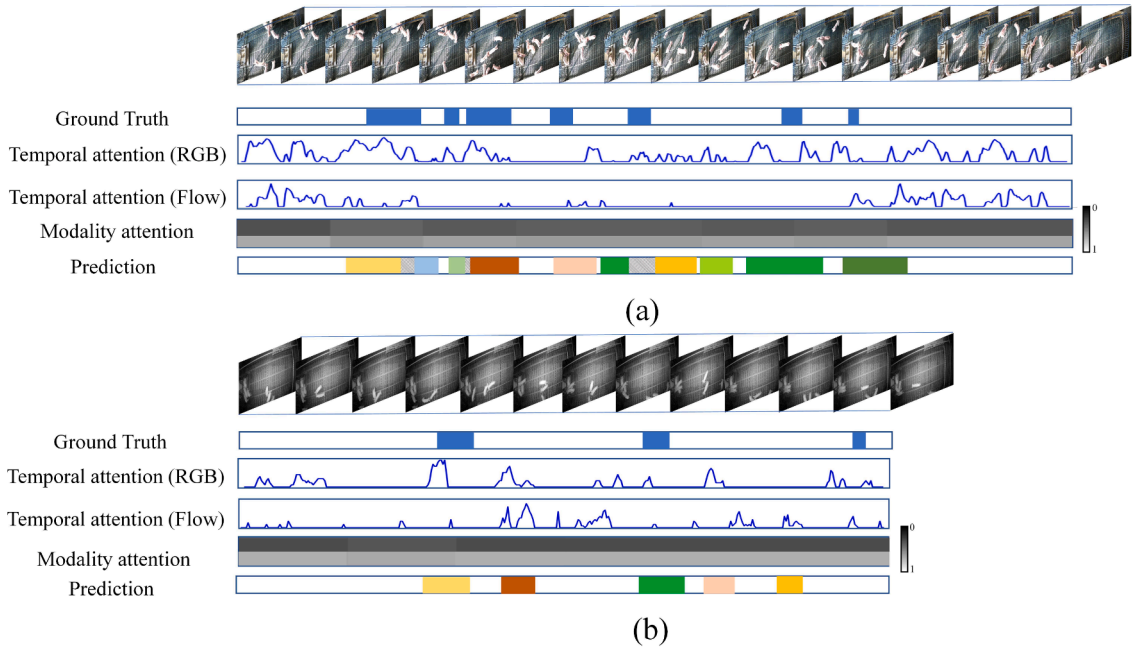

Fig. 11. Weight visualization of different attention block on test set. (a) and (b) represent daytime and night data int the test set.

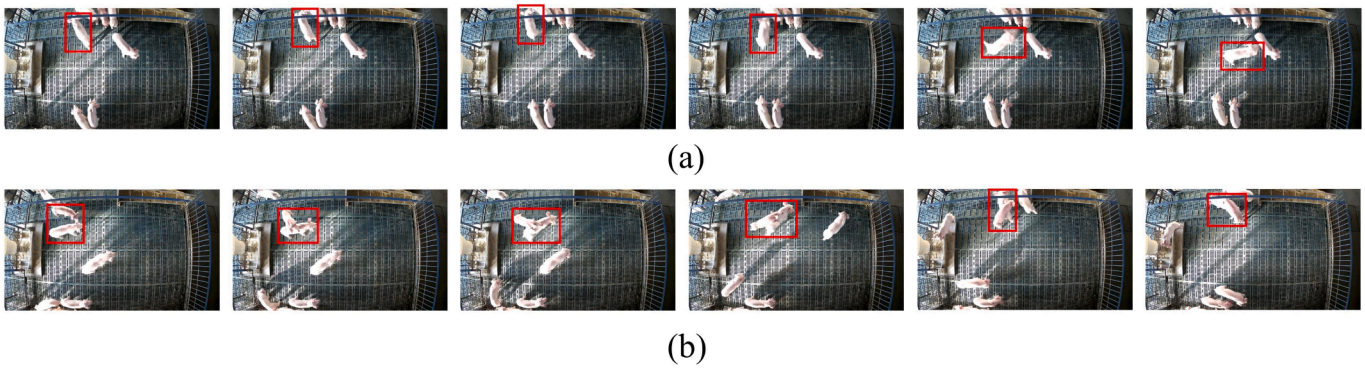

Fig. 12. Partial frames at the start and end in Fig. 9(a). (a) is start frame, and (b) is end frame.

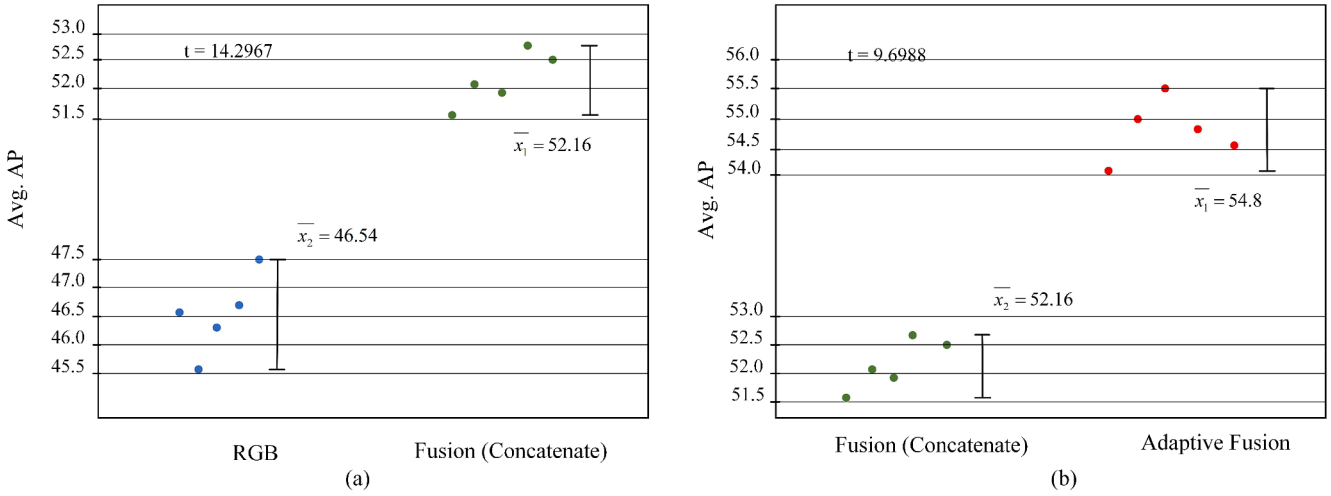

Fig. 13. Performance analysis between RGB model (shown in blue), fusion (concatenate) model (shown in green) and adaptive fusion model (shown in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 8
The effect of the proposed adaptive dual-modality fusion module for aggression temporal proposal generation.

Method	AP(%)@tIoU					Avg. (%)
	0.3	0.4	0.5	0.6	0.7	
Lin et al. (2021)	67.3	62.4	55.5	43.7	31.1	52.0
Lin et al. (2021) + Adaptive fusion	68.4	63.6	56.3	44.2	31.5	52.8
Improvement	+1.1	+1.2	+0.8	+0.5	+0.4	+0.8
Zhang et al. (2022)	82.1	77.8	71.0	59.4	43.9	66.8
Zhang et al. (2022) + Adaptive fusion	82.7	78.5	70.6	59.7	44.0	67.1
Improvement	+0.6	+0.7	−0.4	+0.3	+0.1	+0.3

is employed to compress the temporal dimension and obtain multi-scale features. For each level of pyramid features, a regression head and a classification head are used for coarse prediction. These heads generate the start and end times of aggressive behavior along with its confidence score.

Through the refined prediction block, the most salient boundary features are identified from each coarse prediction proposal using boundary pooling. These features are then used to refine the pyramid features. Based on the refined features, the model further predicts aggressive behavior proposal and outputs the confidence score to indicate the quality of the prediction.

3.5. Loss function

In this study, overall loss function consists of three parts: the regression loss L_{reg} that constrains the start and end frame positions, the classification loss L_{cls} that discriminates whether it is an aggressive behavior, and the IoU (Intersection over Union) loss L_{iou} that constrains the tIoU threshold between the proposal and the ground truth.

$$L = L_{reg} + L_{cls} + L_{iou}\# \tag{7}$$

Classification and regression losses are used twice, once to generate initial proposals and once to refine proposals. Since they are calculated using the same method, we will discuss one of them here. Regression losses include start and end frame localization loss L_{start}, L_{end}, both of them use $L1loss$. $L1loss$ is used to evaluate the distance between model predictions and real frame position. $f(x_i)$ is the predicted value obtained after the model. y_i is the real frame position.

$$L_{start}, L_{end} = \frac{1}{N}\sum_{i=1}^{N} |y_i - f(x_i)|\# \tag{8}$$

The classification loss is evaluated by focal loss. Focal loss is a commonly used method to address the problem of unbalanced data during training. For example, there are typically more background frames than frames with aggressive behavior in video data. Focal loss has been widely used in object detection to address this problem. $f(x_i)$ is the class of the prediction. y_i is the real class.

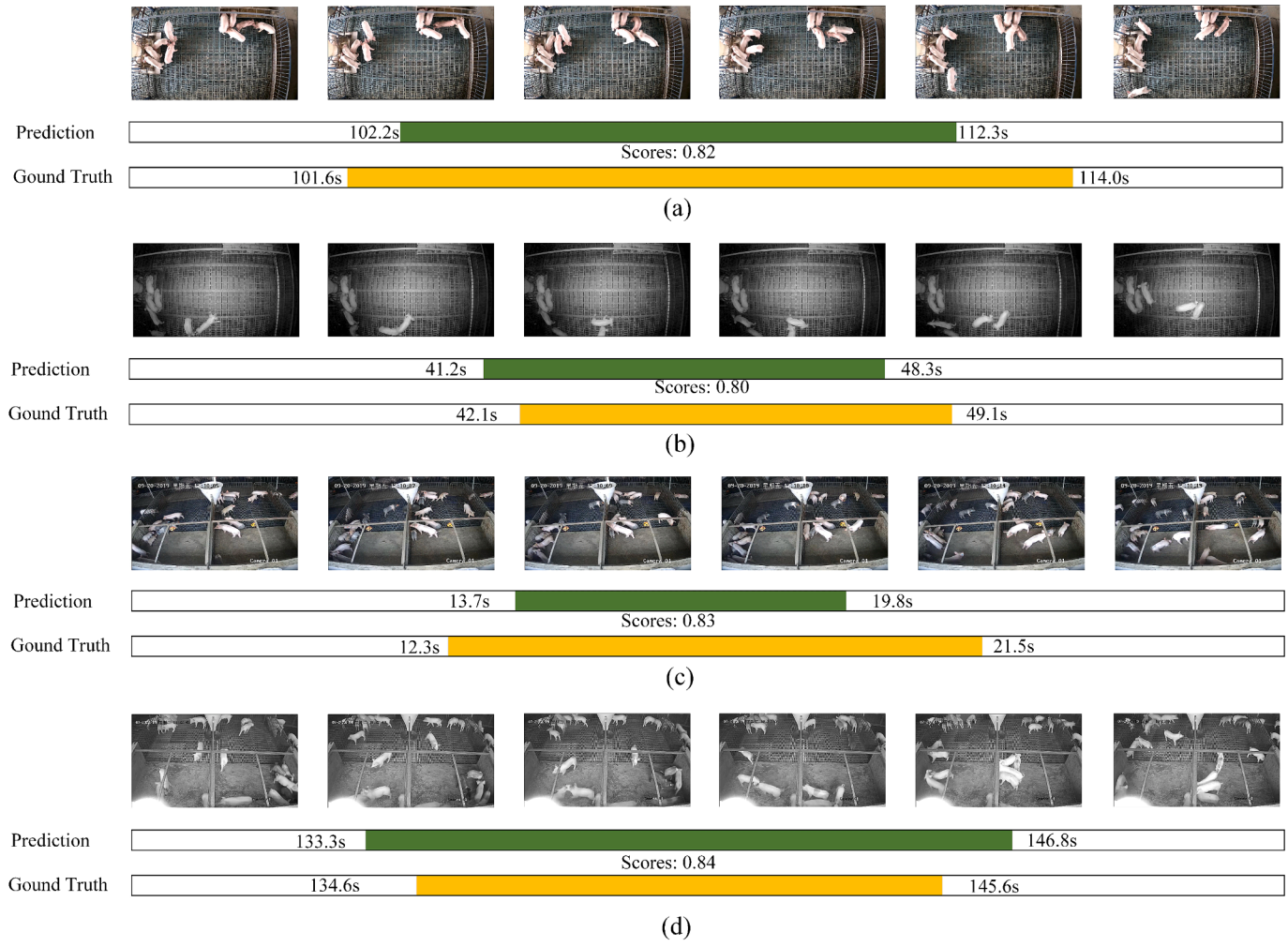

Fig. 14. Detection results in in the different environments. (a) is the detection result in day time recorded from overhead view. (b) is the detection result in night time collected from overhead view. (c) is the detection result in day time recorded from oblique view. (d) is the detection result in night time collected from oblique view.

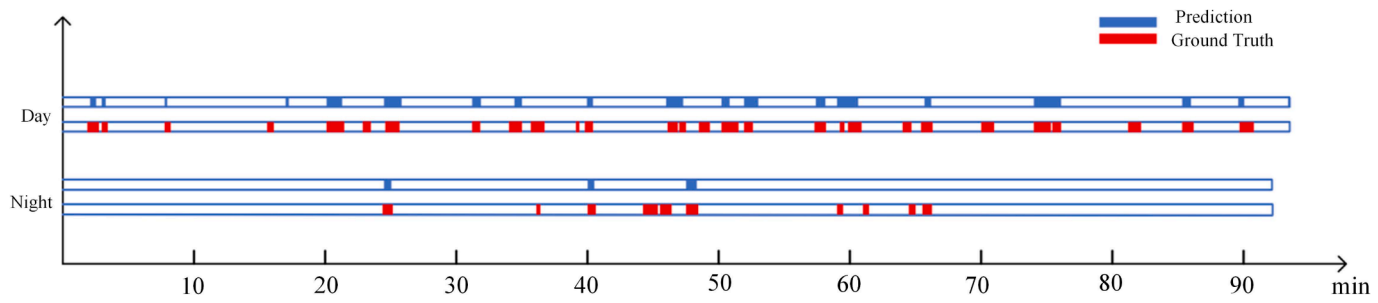

Fig. 15. Aggression temporal behavior detection results on untrimmed group-housed pig surveillance videos.

Table 9
The time consumption for detecting aggressive behavior on a 92-minute video.

Operation	Time
Optical flow estimation	531.6 s
Inference	3.5 s

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} l_{focal}(f(x_i), y_i) \tag{9}$$

In our study, the IoU loss is used to evaluate the completeness of proposals. TIoU is represented as shown in Fig. 8. T_p is the proposal, and T_g is the ground truth. The specific IoU loss function is shown as follow.

$$L_{iou} = -In\frac{\left|T_p \cap T_g\right|}{\left|T_p \cup T_g\right|}\# \tag{10}$$

4. Results and analysis

4.1. Experimental settings

The testing of temporal aggressive behavior detection is conducted on a Python 3.6 platform using PyCharm as the programming software and PyTorch as the deep learning framework. The experiments are conducted on Windows system, and the parameters of the experimental environment are shown in Table 4.

The experiment used Adam optimization as gradient descent algorithm with a learning rate of 1e-5. Learning rate decay was used with a weight of 1e-3. Proposals with confidence score greater than 0.1 were selected during test, other proposals were suppressed. The non-maximum suppression (NMS) algorithm was used to process the proposals, with a parameter setting of 0.5.

4.2. Evaluation metrics

Temporal aggressive behavior detection aims to generate high quality proposals which cover behavior instances with high Average Precision (AP) and high Average Recall (AR). AP and AR under multiple tIoU thresholds are the major evaluation metrics used to assess our method. This involves metrics such as precision P and recall R.

$$R = \frac{TP}{TP + FN}\# \tag{11}$$

$$P = \frac{TP}{TP + FP}\# \tag{12}$$

where TP represents the amount of data for which an aggressive behavior is correctly recognized as an aggressive behavior; FP represents the amount of data incorrectly recognized as aggressive behavior; FN represents the amount of data for which aggressive behavior was incorrectly recognized as non-aggressive behavior.

For the N instances of aggressive behavior, all predicted proposals were sorted by confidence scores. AP combines recall and precision, which is the area under curve with recall as the horizontal coordinate and precision as the vertical coordinate as the number of proposals increases. AP is calculated as shown in Eq. (13).

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

AR is average recall under different tIoU thresholds. AR is calculated as shown in Eq. (14).

$$AR = \frac{1}{N_{tIoU}}\sum_{i=0.1}^{tIoU} r_i\# \tag{14}$$

In addition, we calculate AR under different Average Number of proposals (AN) as AR@AN.

Fig. 16. Limitations analysis of the proposed method in test set. It demonstrates the limitations for long temporal prediction.

4.3. Analysis of detection results and performance evaluation

4.3.1. Ablation experiment of dual-modality fusion model

To evaluate the effectiveness of different model on temporal aggressive behavior detection, we performed comparative experiment. Table 4 and Table 5 show AP values of RGB model, Flow model, concatenating fusion model and adaptive fusion model at different tIoU thresholds, as well as the AR@AN value.

As shown in Table 5, Flow model has a clear advantage over RGB model, with an average AP increase of 2.7. In addition, AP value of Flow model is higher than that of RGB model at all tIoU thresholds. This indicates that motion information is an indispensable factor in the detection of aggressive behavior. The Fusion model, which simply concatenates RGB feature and Flow feature, achieves a 5.9 % increase in average AP over the RGB model and a 3.2 % increase over the Flow model. This result shows that there is complementarity between spatial information extracted from RGB modality and motion information extracted from flow modality. The fusion of different modalities can improve the detection accuracy of aggressive behavior. Adaptive fusion model considers assigning weights to different modalities for each behavior. For example, spatial information may not be distinct enough when detecting biting behavior, a higher flow modality weight should be assigned relative to other behaviors. In addition, temporal attention is used to enhance key frames in long temporal series. As a result, the adaptive fusion module results in a 3 % increase in AP compared to the simple concatenating model.

According to Table 6, it can be seen that adaptive fusion model has an improved performance compared to RGB and Flow model when average number of proposals is 5, 10, and 100. Looking at the dataset statistics in Section 2.3, 51.0% of the video clips contain 2–4 aggressive behaviors. Therefore, we choose AR@5 and AR@10 as the main analysis metrics.

Fig. 9 shows the visualization of different modalities in the same aggressive behavior instance. The first and last frames in the figure represent the frames before and after the aggressive behavior, and the 6 frames in the middle represent the aggressive behavior instance. The visualization of the optical flow color coding is shown in the last column, where the hue represents the motion direction and the saturation represents the motion intensity. Fig. 9(a) shows multiple motion behaviors occurring simultaneously in a large displacement aggressive behavior scene. Only relying on the optical flow modality for detection faces a greater challenge, while combining the spatial information provided by the RGB modality can lead to better detection results. After the stage of aggressive behavior, the movements will still occur on the aggressive moving pigs, which makes the behavior detection based on RGB modality alone difficult. Combining optical flow modality can determine the boundary information of the aggressive behavior more precisely.

Fig. 9(b) depicts a segment of aggressive behavior which happened in dense pig herd. The optical flow modality can observe the motion information in the aggressive behavior instances. As for the RGB modality, the aggressive behavior in the dense pig herd lacks obvious spatial location change information to make a good detection. The detection results of the RGB model are shown in Fig. 9(b). For this type of behavior, the detector relies more on optical flow data.

Fig. 10 demonstrates the effectiveness of our adaptive fusion model using two examples from the test set. In Fig. 10(a), both the RGB and Flow models fail to predict the aggressive behavior with accurate boundaries, while our adaptive fusion model combines information from both modalities, achieves higher tIoU thresholds and confidence scores. In Fig. 10(b), the RGB model performs poorly in detecting the aggressive behavior in the dense pig herd. The flow model is more prone to misjudging certain movements that occur before the start frame. The adaptive fusion model achieves better results by combining the different information of dual-modality.

4.3.2. Ablation experiment of different attention block

To demonstrate the effectiveness of modality attention and temporal attention in the adaptive fusion model, evaluation experiments were conducted. Table 6 shows the improvement of each block on the overall model, where we mainly compared the average ap, ap at 0.3 threshold, ap at 0.5 threshold, and AR@10.

Fusion (concatenate): The basic model for detecting aggression behavior in group-housed pigs based on concatenating fusion. In the model, a pre-trained I3D network is used to extract features from the dual-modality data. The temporal pyramid network is overlaid on the input dual-modality data, and 1D convolution is used to obtain aggressive behavior proposals.

Fusion (modality attention): Based on the fusion (concatenate) model, modality attention is used to adaptively obtain the weight information of different modalities within the segment.

Fusion (temporal attention): Based on the fusion (concatenate) model, temporal attention is used to enhance key frames for both modalities.

Fusion (both attention): Based on the fusion (concatenate) model, modality attention and temporal attention are added.

From the Table 7, we can see that the original model was improved by adding modality attention module or temporal attention module. The fusion of the two attention modules has further improved the performance. This shows that the two modules improved the model in different ways. The fusion (both attention) model showed a more significant improvement compared to the fusion (concatenate) model, proving that the adaptive dual-modality fusion module is effective.

Fig. 11 shows the results on the test set, with a 12-second interval between each frame. Fig. 11(a) is from daytime, with a video duration of 212 s and 10 proposals. Fig. 11(b) is from the night, with a video duration of 144 s and 5 proposals. The frequency of aggressive behavior is lower at night. The first and last rows respectively represent the ground truth action instances and the predicted proposals. The second and third respectively row show the temporal attention weights of the RGB and Flow modalities. The line graphs represent the changing of the attention weights over the entire temporal sequence. The fourth row represents the modality attention weights in the video segment. The top and bottom represent the weight values of the RGB and optical flow modality data. Optical flow tends with a higher weight in the detection of aggressive behavior, but weight values is also changed on different clips. From Fig. 11, it shows that temporal attention achieved the purpose of automatically capturing important behavioral temporal features and different modalities focus on different clips. The combination of RGB and flow modalities can produce better results.

The beginning of Fig. 11(a) is shown in Fig. 12(a). During this time, the pigs in the upper left corner had strong motion information and changes in spatial position, but did not interact with other pigs. In our classification, this behavior is a play behavior. Due to the similarity between play behavior and aggressive behavior, this may be the reason why the RGB and flow modality temporal attention weight values are relatively high during this period. Fig. 12(b) represents the end of Fig. 11 (a), which is a period of mounting behavior between pigs, and the pig being climbed will have some change in spatial position, but it is not an aggressive behavior. These two segments have lower confidence in subsequent networks, and use the NMS algorithm to remove the proposal after generating the proposal by confidence.

The *t*-test method is used to verify the significance of model improvement. The principle of the *t*-test is to calculate the t-value using the means and standard deviations of two samples, and then obtain the p-value through the t-value. By comparing the p-value with a preset significance level (such as 0.05), if the p-value is less than the significance level, it is considered that there is a significant difference between the two samples.

For the *t*-test of two independent samples, the formula is as follows:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \# \quad (15)$$

Where x1 and x2 represent the means of the two experiments, n1 and n2 represent the number of experiments, and s1 and s2 represent the standard deviations of the two experiments.

To verify the robust of proposal method, we carry out each experiment 5 times and report the performance in Fig. 13. Since a big t-value corresponds to a very small p-value, it is confident to conclude that proposal method can consistently improve the performance of aggressive behavior.

4.3.3. Effect verification of adaptive fusion module on THUMOS14 dataset

THUMOS14 is a public dataset for temporal action detection containing dense actions, with 200 training videos and 212 test videos spanning 20 different categories. On average, each video contains 15 action instances.

Table 8 reports the effect of the proposed adaptive dual-modality fusion module on the THUMOS14 dataset. Under the average AP, the methods of Lin et al. (2021) and Zhang et al. (2022) receive 0.8 % and 0.3 % improvements. Lin et al. (2021) trains RGB and optical flow models separately to generate proposals, and finally achieves fusion based on confidence scores. Zhang et al. (2022) integrates global temporal information based on transformer structure. The adaptive dual-modality fusion module can assign different weights to different video segments for the improvement of these detection performance. Zhang et al. (2022) uses self-attention mechanism to increase the temporal receptive field, but it may increase the similarity between the background frame and the behavior frame when applied to the detection of aggressive behavior. In aggressive behavior detection dataset, 89.2 % of aggression lasted less than 20 s, Zhang's method will be a primary consideration when locating long aggressive behavior in the future.

4.4. Generalization validation of aggression detection

In order to evaluate the verify the generalization of aggression detection method in different environments, the corresponding experiments were carried out in different views during day time and night time. As shown in the Fig. 14, aggressive behavior in different environments can be located which proves that the proposed method has generalization ability.

To evaluate the detection ability in untrimmed surveillance video, the experiment was conducted on day and night surveillance video, with durations of 95 min and 92 min. The experiment results are shown in Fig. 15. For the day time video, the top 30 proposals with the highest confidence were selected, and for the night time video, the top 10 proposals with the highest confidence were selected. As shown in Fig. 15, the predicted proposals can cover almost all the real instances.

In the detection of the 92-minute surveillance video, optical flow estimation using the TV-L1 algorithm took 531.6 s. The inference including feature extraction, adaptive dual-modality fusion and temporal proposal generation took 3.5 s. The time consumption is shown in Table 9.

5. Analysis of method limitations

Fig. 16 shows some examples of detection failures. The predicted examples are all from the top ten high confidence proposals of each video. In Fig. 16, an aggressive behavior lasting 58.8 s with a correct label of [51.2, 110.0] was not fully detected. We selected the top ten predictions of this segment, among which six segments are located within the label: [51.8, 63.3], [52.8, 67.2], [70.4, 82.3], [77.2, 87.9], [98.4, 108.9], and [102.6, 111.0]. The longest predicted length is 15 s.

This method tends to predict shorter proposals, possibly due to the limitation of the convolutional structure's inability to achieve a global receptive field and the restriction on proposal size caused by the model's sliding window size. In subsequent work, self-attention structure should be considered to enhance the temporal receptive field, or employing gating mechanisms to memorize relationships between different segments. IoU loss can be further improved, similar to advancements made in the field of object detection. Additionally, this paper utilizes additional optical flow modality data as input. Optical flow data needs to be manually obtained offline which prevents end-to-end training. This affects the subsequent deployment and production application of the model. In future work, the integration of deep learning-based optical flow generation algorithms with the proposed method should be considered to achieve end-to-end training and application.

6. Conclusion

In this study, temporal behavior detection method was proposed to automatically determine the occurring and temporal interval of aggressive behavior from surveillance video. To extract aggressive behavior pattern, an adaptive dual-modality fusion module was designed to fuse spatial and motion information extracted by a pre-trained I3D model from RGB and Flow modalities. The proposed method achieved an AP value of 68.0 %, an AR value of 77.8 % in average number of proposals at 100. The effectiveness of this method was verified in raw surveillance video. The results showed the feasibility of using temporal behavior detection in behavior monitor and the effectiveness of extracting aggressive behavior pattern by adaptively fusing spatial and motion information. In the future, different seasons and perspectives data will be added to improve the robustness of the proposed method. In addition, the temporal behavior detection method can be extended to other important behaviors, such as feeding, drinking, and nursing. Multi-behavior detection can provide a more comprehensive understanding of pigs' activities and interactions. Meanwhile, this method can not only be used for piglets but also for monitoring other types of animals, such as cows, sheep, or chickens, to facilitate intelligent and automated farming. For different animals and complex environments, transfer learning is a good approach to achieve migration between different scenarios. To enhance generalizability in video analysis of different types of animals and environments, a generative large pre-trained model needs to be constructed.

Author contribution

Kai Yan: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. **Baisheng Dai:** Data curation, Formal analysis, Funding acquisition, Methodology Writing – review & editing. **Honggui Liu:** Resources, Validation. **Yanling Yin:** Funding acquisition, Software. **Xiao Li:** Resources, Validation. **Renbiao Wu:** Resources, Software, Validation. **Weizheng Shen:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

CRediT authorship contribution statement

Kai Yan: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Baisheng Dai:** Writing – review & editing, Methodology, Funding acquisition, Formal analysis, Data curation. **Honggui Liu:** Validation, Resources. **Yanling Yin:** Software, Funding acquisition. **Xiao Li:** Validation, Resources. **Renbiao Wu:** Validation, Software, Resources. **Weizheng Shen:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

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

Data availability

I have shared the link to my data

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 31902210 and 32172784, in part by the Natural Science Foundation of Heilongjiang Province under Grant QC2018074 and YQ2023C012.

References

Bati, C.T., Ser, G., 2023. SHEEPFEARNET: Sheep fear test behaviors classification approach from video data based on optical flow and convolutional neural networks [J]. Comput. Electron. Agric. 204, 107540.

Carreira J., Zisserman A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 6299-6308.

Chen, C., Zhu, W., Ma, C., et al., 2017. Image motion feature extraction for recognition of aggressive behaviors among group-housed pigs[J]. Comput. Electron. Agric. 142, 380–387.

Chen, C., Zhu, W., Guo, Y., et al., 2018. A kinetic energy model based on machine vision for recognition of aggressive behaviours among group-housed pigs[J]. Livest. Sci. 218, 70–78.

Chen, C., Zhu, W., Liu, D., et al., 2019. Detection of aggressive behaviours in pigs using a RealSence depth sensor[J]. Comput. Electron. Agric. 166, 105003.

Chen, C., Zhu, W., Steibel, J., et al., 2020. Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory[J]. Comput. Electron. Agric. 169, 105166.

Coutellier, L., Arnould, C., Boissy, A., et al., 2007. Pig's responses to repeated social regrouping and relocation during the growing-finishing period[J]. Appl. Anim. Behav. Sci. 105 (1–3), 102–114.

Dawkins, M.S., Lee, H., Waitt, C.D., et al., 2009. Optical flow patterns in broiler chicken flocks as automated measures of behaviour and gait[J]. Appl. Anim. Behav. Sci. 119 (3–4), 203–209.

de Groot, J., Ruis, M.A.W., Scholten, J.W., et al., 2001. Long-term effects of social stress on antiviral immunity in pigs[J]. Physiol. Behav. 73 (1–2), 145–158.

Gan, H., Ou, M., Huang, E., et al., 2021a. Automated detection and analysis of social behaviors among preweaning piglets using key point-based spatial and temporal features[J]. Comput. Electron. Agric. 188, 106357.

Gan, H., Li, S., Ou, M., et al., 2021b. Fast and accurate detection of lactating sow nursing behavior with CNN-based optical flow and features[J]. Comput. Electron. Agric. 189, 106384.

Gao, Y., Chen, B., Liao, H.M., et al., 2019. Recognition method for aggressive behavior of group pigs based on deep learning[J]. Transactions of the Chinese Society of Agricultural Engineering. 35 (23), 192–200.

Gao, Y., Yan, K., Dai, B., et al., 2023. Recognition of aggressive behavior of group-housed pigs based on CNN-GRU hybrid model with spatio-temporal attention mechanism [J]. Comput. Electron. Agric. 205, 107606.

Gonyou, H.W., 2001. The social behaviour of pigs[M]//Social behaviour in farm animals. CABI publishing, Wallingford UK, pp. 147–176.

Gronskyte, R., Clemmensen, L.H., Hviid, M.S., et al., 2016. Monitoring pig movement at the slaughterhouse using optical flow and modified angular histograms[J]. Biosyst. Eng. 141, 19–30.

He, D.J., Liu, D., Zhao, K.X., 2016. Review of perceiving animal information and behavior in precision livestock farming[J]. Transactions of the Chinese Society Agricultural Machinery. 47 (5), 231–244.

Kongsted, A.G., 2004. Stress and fear as possible mediators of reproduction problems in group housed sows: a review[J]. Acta Agriculturae Scandinavica, Section A-Animal Science. 54 (2), 58–66.

Lee, J., Jin, L., Park, D., et al., 2016. Automatic recognition of aggressive behavior in pigs using a kinect depth sensor[J]. Sensors 16 (5), 631.

Lin T., Liu X., Li X., et al., 2019. Bmn: Boundary-matching network for temporal action proposal generation[C]//Proceedings of the IEEE/CVF international conference on computer vision. 3889-3898.

Lin C., Xu C., Luo D., et al., 2021. Learning salient boundary feature for anchor-free temporal action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 3320-3329.

Liu, Q., Wang, Z., 2020. Progressive Boundary Refinement Network for Temporal Action Detection[c]//proceedings of the AAAI Conference on Artificial Intelligence. 34 (07), 11612–11619.

McGlone, J.J., 1985. A quantitative ethogram of aggressive and submissive behaviors in recently regrouped pigs[J]. J. Anim. Sci. 61 (3), 556–566.

Oczak, M., Viazzi, S., Ismayilova, G., et al., 2014. Classification of aggressive behaviour in pigs by activity index and multilayer feed forward neural network[J]. Biosyst. Eng. 119, 89–97.

Perez M., Toler-Franklin C., 2023. CNN-Based Action Recognition and Pose Estimation for Classifying Animal Behavior from Videos: A Survey[J]. arXiv preprint arXiv: 2301.06187.

Shi D. , Zhong Y. , Cao Q. , et al., 2023. Tridet: Temporal action detection with relative boundary modeling [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 18857-18866.

Spoolder, H.A.M., Edwards, S.A., Corning, S., 2000. Aggression among finishing pigs following mixing in kennelled and unkennelled accommodation[J]. Livest. Prod. Sci. 63 (2), 121–129.

Su, H., Gan, W., Wu, W., et al., 2021. Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation[C]// Proceedings of the. AAAI Conference on Artificial Intelligence. 35 (3), 2602–2610.

Turner, S.P., Farnworth, M.J., White, I.M.S., et al., 2006. The accumulation of skin lesions and their use as a predictor of individual aggressiveness in pigs[J]. Appl. Anim. Behav. Sci. 96 (3–4), 245–259.

Viazzi, S., Ismayilova, G., Oczak, M., et al., 2014. Image feature extraction for classification of aggressive interactions among pigs[J]. Comput. Electron. Agric. 104, 57–62.

Wang W., Tran D., Feiszli M., 2020. What makes training multi-modal classification networks hard? [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 12695-12705.

Xu H., Das A., Saenko K., 2017. R-c3d: Region convolutional 3d network for temporal activity detection[C]//Proceedings of the IEEE international conference on computer vision. 5783-5792.

Xu M., Zhao C., Rojas D.S., et al., 2020. G-tad: Sub-graph localization for temporal action detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 10156-10165.

Yang, L., Han, J., Zhao, T., et al., 2022. Structured attention composition for temporal action localization[J]. IEEE Transactions on Image Processing. Early Access. https://doi.org/10.1109/TIP.2022.3180925.

Yang, A., Huang, H., Yang, X., et al., 2019. Automated video analysis of sow nursing behavior based on fully convolutional network and oriented optical flow[J]. Comput. Electron. Agric. 167, 105048.

Yang, A., Huang, H., Zheng, B., et al., 2020. An automatic recognition framework for sow daily behaviours based on motion and image analyses[J]. Biosyst. Eng. 192, 56–71.

Yoon, D.H., Cho, N.G., Lee, S.W., 2020. A novel online action detection framework from untrimmed video streams[J]. Pattern Recogn.. 106, 107396.

Yu J., Hong J., 2021. Sarnet: self-attention assisted ranking network for temporal action proposal generation[C]//2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 1062-1067.

Zhang, C.L., Wu, J., Li, Y., 2022. Actionformer: Localizing moments of actions with transformers[C]//European Conference on Computer Vision. Springer Nature Switzerland, Cham, pp. 492–510.