Computers and Electronics in Agriculture 205 (2023) 107606

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

# Recognition of aggressive behavior of group-housed pigs based on CNN-GRU hybrid model with spatio-temporal attention mechanism

Yue Gao [a], Kai Yan [a], Baisheng Dai [a,b,*], Hongmin Sun [a,**], Yanling Yin [a,b], Runze Liu [c], 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

ARTICLE INFO

*Keywords:*
Aggressive behavior recognition
Attention mechanism
CNN
GRU
Group-housed pigs

ABSTRACT

Aggressive behavior of group-housed pigs seriously affects farm economy and animal welfare. Automatic and accurate recognition of aggressive behavior of group-housed pigs is thus important for farm production management. This study proposes a hybrid model that combines convolutional neural network (CNN) and gated recurrent unit (GRU) to differentiate aggressive and other behaviors from surveillance videos. The CNN network served as a spatial feature extractor to learn appearance representations of behavior in each individual frame, while the GRU network served as a temporal feature extractor to learn motion representations of behavior in a behavior episode. More importantly, to focus on the saliency features in both the spatial domain and the temporal domain of behavior, a specific spatio-temporal attention mechanism was designed and integrated in the CNN-GRU hybrid model to improve the effect of aggressive behavior recognition. To evaluate the proposed model, a behavior video dataset consisting of 5530 behavior episodes about 10 piglets. The accuracy of the proposed hybrid model conducted on the test set were 94.8 %. The results showed that the proposed hybrid model integrated with spatio-temporal attention performed better than the model with an independent spatial or temporal attention and the model without attention mechanism, and achieved a competitive performance of aggressive behavior recognition over the state-of-the-art approaches. We shared our behavior video dataset at https://github.com/IPCLab-NEAU/Aggressive-Behavior-Recognition for precision livestock farming research community.

## 1. Introduction

Modern pig breeding is becoming more and more intensive and large-scale. Pigs housed in groups may experience multiple mixed groups during the management of specific farms, and there is a high probability of aggressive behavior after mixing (Buettner et al., 2015). Aggressive behavior can seriously affect growth, reproduction rates and welfare of pigs (Kongsted, 2004), therefore, in the pig production, it is necessary to pay attention to whether pigs participate in aggressive behavior, and then further judge whether human intervention is required (He et al., 2016). With the development of computer vision and deep learning, some scholars were advocating the use of automatic and non-intrusive aggressive behavior recognition methods. Compared to traditional manual observation, vision-based methods are efficient in real time. To a certain extent, it can ensure the welfare of pigs, reduce labor costs, avoid the risk of cross-infection of human and animal diseases (Liu et al., 2014, Neethirajan, 2017, Li et al., 2021), and also facilitate researchers in the field of animal science to observe and research the behavior of animals.

Existing studies on aggressive behavior recognition of group-housed pigs can be divided into two categories: traditional methods and deep learning based methods to recognize aggressive behaviors in videos. In the traditional methods, Viazzi et al. (2014) firstly extracted two motion features of pigs related to aggressive behavior based on motion history images, and used Linear Discriminant Analysis (LDA) to recognize aggressive behaviors. Lee et al. (2016) defined multiple features with respect to the speed and the distance of standing pigs based on depth image data, and then trained support vector machine (SVM) to recognize

* 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), hmsun@neau.edu.cn (H. Sun).

https://doi.org/10.1016/j.compag.2022.107606
Received 19 March 2022; Received in revised form 18 December 2022; Accepted 27 December 2022
Available online 12 January 2023

aggressive behavior of pigs. Dong et al. (2021) proposed an activity index calculation method to recognize whether aggressive behavior occurs, which solves the problems of varieties of breeding facilities and dynamic background environment existing in inter-frame difference method. To further distinguish medium and high aggressive behaviors, Oczak et al. (2014) extracted relevant features of pig herd activity index, and trained a multi-layer feedforward neural network to recognize aggressive behaviors with different degrees of intensity. Chen et al. (2017) separated aggressive pigs by using connection area and adhesion index, and then calculated acceleration features to distinguish medium and high aggressive behaviors. Chen et al. (2018) located contour feature points of pigs in aggressive behavior, and further extracted a kinetic energy of these points. The kinetic energy difference between adjacent frames was then used as a feature to recognize medium and high aggressive behavior. Above-mentioned works mainly employed traditional image processing technology to manually define and extract one or several features of aggressive behavior, and achieved the recognition of aggressive behaviors of group-housed pigs. However, the effectiveness of those hand-crafted features is susceptible to adhesion and occlusion of pigs, poor lighting conditions, rapid body growth of pigs and complex patterns of behavior. The recognition performance of above methods will deteriorate in practical breeding condition.

Deep learning technology has the merit of automatic feature learning, which can remarkably improve the representation of features to distinguish different behaviors (Gao et al., 2019). In recent years, researchers have carried out some studies on recognizing aggressive behavior of group-housed pigs based on deep learning methods. Considering that behaviors appeared different not only in spatial domain but also in temporal domain of video data, existing deep learning methods pursued learning these two aspects of behavior features automatically. Gao et al. (2019) proposed a 3D convolutional neural network integrating multi-scale features fusion to learn both spatial and temporal features and built an end-to-end model for identifying aggressive behaviors. Chen et al. (2020) adopted a spatio-temporal model combining convolutional neural network (CNN) and long short-term memory (LSTM) to recognize aggressive behavior of group-housed pigs. The above two methods learned behavior features from video frame sequences directly, which solved the mentioned issues of hand-crafted features. However, it is worth noting that the local area of pig aggression is small compared to the whole activity scene appeared in video frames, and meanwhile the discriminable motion information over frame sequences generally reflected in several key frames (Pei et al., 2019). Liu et al. (2020) proposed a two-stage strategy of tail-biting behavior recognition. Object detection and tracking methods were first adopted to obtain sub-videos of pairwise interactions of pigs. And then a combined model of CNN and LSTM was used to classify behavior categories of those sub-videos. Extracting features from whole spatial or temporal domain inevitably introduce many redundant information, which will influence the discriminability of features for behavior recognition. To focus on the local area of pig aggressions. However, the recognition performance of this two-stage framework is relatively depending on pig detection and tracking, which also be challenging tasks in group-housed condition.

With the continuous development of attention mechanism, more and more scholars use it in video field for video classification and behavior recognition. Sudhakaran et al. (2019) used a mechanism of Long Short-Term Attention to focus on features in the spatial part while smoothly tracking attention throughout the video sequence, thus enabling more effective action recognition. Zhang et al. (2022) designed a What-Where-When Attention Network (W3AN), which learns temporal-spatial correlations from consecutive video frames through spatial attention and temporal attention modules for better video person re-identification. From the above, it can be seen that the use of attention mechanism can better extract the spatial temporal and other features of the video, which leads to more accurate and effective classification of herd-raised pig fighting behavior. In addition, key frames contribute more to the task of behavior recognition were still be neglected in that framework. Dong et al. (2022) used an attention-based approach to identify key frames in the video for better action recognition.

In this paper, a spatio-temporal attention mechanism is specially designed and integrated in a novel CNN-GRU hybrid model to achieve an automatic recognition of aggressive behavior of group-housed pigs. Among the proposed attention-based model, the spatial attention module devotes to guide VGG16 convolutional neural network (Simonyan and Zisserman, 2014), used as the spatial feature extractor, to focus on the local area where aggression occurs. While the temporal attention module devotes to guide a GRU network (Cho et al., 2014), used as the temporal feature extractor, to pay more attention to the key frames which provided more discriminable motion information. The main contributions of this work are follows: First, a behavior video dataset consisting of aggressive and non-aggressive behaviors was built for model training and testing, which is available at https://github.com/IPCLab-NEAU/Aggressive-Behavior-Recognition. Second, a novel CNN-GRU hybrid model with spatio-temporal attention mechanism was proposed for aggression recognition. Third, the effectiveness of the designed spatio-temporal attention mechanism was proved by comparative experiments, and the proposed hybrid model achieved a competitive performance in recognition of aggressive behavior of group-housed pigs.

## 2. Materials and methods

### 2.1. Data acquisition

The video data was obtained from the Harbin HongFu Pig Farm in October 2018. The pig pen was 4.3 m long and 2.3 m wide. The age of the pigs is 35–42 days old, there is no sex differentiation, and the rearing density is about 10 pigs per pen. There were feeding troughs and nipple drinkers inside the pen. Due to the frequent aggressive behaviors within three days after pigs were mixed group (Spoolder et al., 2000), the video data of about 10 Large white × Landrace piglets within 72 h after mixing together were recorded. In the subsequent data processing, we will automatically split these videos into 3 s video sets, and then annotate the video sets. The Hikvision DS-2CD3345D-I camera was used to record RGB video with a vertical downward angle of view. And 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. All animal experiments should comply with the ARRIVE guidelines and should be carried out in accordance with the U.K. Animals (Scientific Procedures) Act, 1986 and associated guidelines, EU Directive 2010/63/EU for animal experiments, or the National Research Council's Guide for the Care and Use of Laboratory Animals.

### 2.2. Dataset construction

In order to evaluate the recognition performance of pig aggressive behavior, the dataset in this paper needs to include aggressive and non-aggressive behavior video episodes.

Pig aggressive behavior involves the interaction of multiple pigs' states and is a complex, gradual behavior. At the beginning of the aggressive interaction behavior, pigs will generally perform initial probing by sniffing and light pushing, and then as the aggressive behavior intensifies, more intense biting and bumping will generally occur. At the most intense stage of aggressive behavior, ear biting, tail biting, body biting, etc. will occur. Besides, chasing and trampling are also common aggressive behaviors among pigs. Meanwhile, among the aggressive behaviors, biting usually more frequently in aggressive behaviors and tends to cause skin wounds (Kongsted, 2004, Turner et al., 2006, Chen et al., 2020).

Aggressive behaviors of group-housed pigs defined in this dataset mainly include biting, knocking, trample and chasing. Biting can be

further divided into tail biting, ear biting and body biting, and knocking is divided into knocking head and knocking body. Specific descriptions are shown in Table 1, additionally, mounting, playing, lying, feeding and drinking were recognized as non-aggressive behaviors (Yang et al., 2021).

After determining the classification criteria of the dataset, the initial observation of the data mentioned in section 2.1 shows that most of the fighting behaviors only lasted for a short period of time such as tearing and hitting, while aggressive behaviors of shorter duration occurred more frequently, so 3 s was chosen as a standard parameter when cropping the data.

We invite professionals to conduct manual observation of the obtained data, and selects 541 video episodes of aggressive behavior with a duration of 3 s, and correspondingly selects 565 video episodes of non-aggressive behavior with a duration of 3 s. Fig. 1 shows comparison of the spatial representation of aggressive behaviors in our dataset.

All video data are divided into training set, validation set and test set according to the ratio of 6:2:2. The data are randomly divided to form the dataset used in this study, then data augmentation is performed as in Fig. 2. The data's distribution is shown in Table 2.

### 2.3. Aggressive behavior recognition model for group-housed pigs

The main workflow of the proposed hybrid model is summarized in Fig. 3. The frame sequences of behavior video episode were inputted to the proposed model for learning behavior features. To learn both spatial and temporal features, the proposed model consisted of two fe5ature extractors, i.e., spatial feature extractor and temporal feature extractor. Spatial feature extractor integrated with a spatial attention mechanism extracted appearance representations of behavior from each individual frame, while temporal feature extractor integrating with a temporal attention mechanism extracted motion representations of behavior from spatial features over different frames. Based on these learned features, aggressive and non-aggressive behaviors can be recognized.

#### 2.3.1. Spatial feature extractor

The spatial feature extractor in this study consists of the conv-block of VGG16 model and a spatial attention module. The conv-block can extract the spatial appearance features of the frame; the role of the spatial attention module is to improve the feature expression of local areas, thereby enhancing the feature expression of the area where aggressive behavior occurs, and weakening the feature expression of irrelevant areas. Fig. 4 shows the complete architecture of the spatial feature extractor.

**Table 1**
Definition of aggressive behavior.

| Behavior classification | Behavior name | Behavior description | |
|---|---|---|---|
| Aggression | Bite | 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) |
| | Hit | 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 | |
| non-aggression | Chasing | Chase between parties after aggressive behavior | |
| | Mounting | initial pig gradually approaches mounted pig and placing its foreleg span on it (This riding behavior lasts for a longer period of time) | |
| | Playing | Play by itself or with other pigs (Activity is similar to aggression, but less vigorous) | |
| | Lying | Lie down in the resting area | |
| | Feeding | Feed in the eating trough | |
| | Drinking | Drink under the waterer | |

In this study, the spatial attention module is placed after the fifth convolutional block of VGG16, and the advantages of adding spatial attention module at this position to extract the spatially saliency features of each frame are shown in Section 2.3. The spatial attention module consists of three convolutional layers and the $1 \times 1 \times 1$ parameter of a 2D locally connected layer.

For each frame, the input frame with size $224 \times 224 \times 3$ is filtered by VGG16 conv-blocks.

Then the spatial attention module takes as input the output of the fifth conv-block and extracts the salient regions of the frame. The generated attention weight containing salience is filtered by the third convolutional layer that has 512 kernels of size $1 \times 1$. The formula description of the weight modeling process of the attention module is shown in formula (1). The feature information x is the input of the attention model. After the function transformation corresponding to spatial attention module, the attention weight Att(x) for x is obtained, and then multiply the learned weight Att(x) with the original feature information x to complete the screening of the original features, generate spatial features (Song et al., 2021), and finally flatten the features through flatten layer to obtain a 25088-dimensional spatial feature vector. The spatial feature vector is used as the input of temporal feature extractor.

$$Att(x) = SpatialAttention(x) \tag{1}$$

$$X = Att(x) \odot x \tag{2}$$

In the weight sharing strategy, the 2D local connection layer in this module is different from the ordinary convolutional layer. The interlayer neurons are only connected to the local range in a fully connected manner, and the neurons that exceed the local range are not connected, and different connections are made. There are independent parameters between the relations, which reduces the connection outside the receptive field. This method is more suitable for the case where the data features have different distributions in different regions in this study.

#### 2.3.2. Temporal feature extractor

In temporal part, this study uses the Gated Recurrent Unit (GRU) as the backbone network of the temporal feature extractor. The GRU network receives the input $x_t$ at the current moment and the state at the previous moment $h_{t-1}$, gets the current state $h_t$ of the network and use it as the input for the next moment. The schematic diagram of the structure of the GRU is shown in Fig. 5.

In Fig. 5, $\oplus$ is the addition operation,$\otimes$ is Hadamard product,$\sigma$ is the sigmoid function, *tanh* is tanh function. At the t time step, the update rules of update gate z, reset gate r, current memory features $h_t^{'}$ and memory features of the output of the t gated recurrent unit $h_t$ are as follows:

$$z_t = sigmoid(w_z x_t + u_z h_{t-1}) \tag{3}$$

$$r_t = sigmoid(w_r x_t + u_r h_{t-1}) \tag{4}$$

$$h_t^{'} = \tanh(w x_t + r_t \otimes u h_{t-1}) \tag{5}$$

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes h_t^{'} \tag{6}$$

In an episode, each frame contains different information. Some frames contain features that contribute significantly to the behavior recognition, while some single-frame images may only contain redundant information. In order to be able to focus on some key frames in the sequence, reduce the interference of redundant frames, and ensure a stable recognition effect in this study, we choose to combine temporal attention in GRU to solve the above problems. The model can automatically find and locate key frames and assign larger weights to them, reducing the attention to redundant frames, and can ensure a better

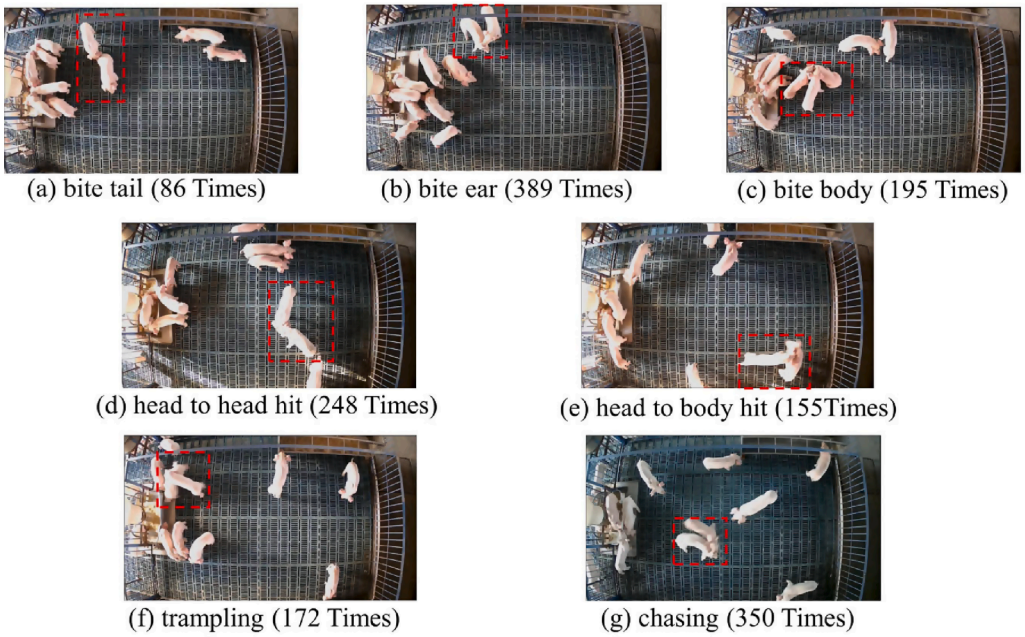


**Fig. 1.** Comparison of data on aggressive behaviors.

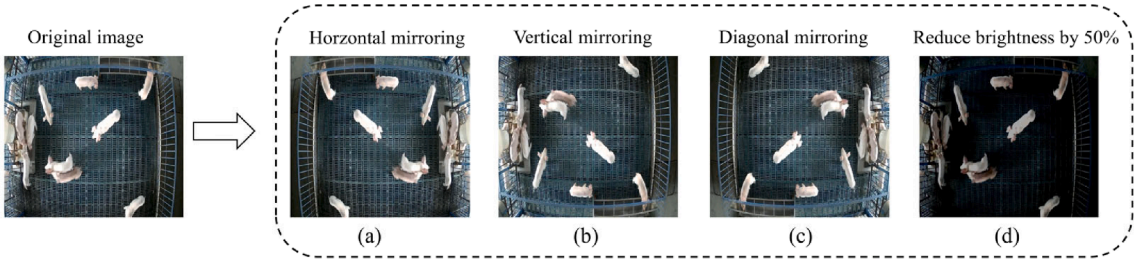


**Fig. 2.** Use data augmentation to meet the data requirement of training model and increase the diversity of data. The methods used for data augmentation are: (a) horizontal mirroring (b) vertical mirroring (c) diagonal mirroring (d) reduce brightness by 50 %.

**Table 2**
Allocation of aggressive and non-aggressive episodes of group-housed pigs.

| Dataset category | Behavior classes | Number (original dataset) | Number (augmented dataset) |
|---|---|---|---|
| Train dataset | Aggression | 324 | 1620 |
| | Non-aggression | 340 | 1700 |
| Test dataset | Aggression | 108 | 540 |
| | Non-aggression | 113 | 565 |
| Validation dataset | Aggression | 108 | 540 |
| | Non-aggression | 113 | 565 |

recognition appearance without manual data selected. The input of temporal feature extractor is the feature vector through spatial feature extractor, and then temporal feature extractor is used to model the temporal relationship between these feature vectors. So as to prepare for the task of group-housed pigs aggressive behavior recognition. Schematic diagram of the temporal feature extractor as shown in Fig. 6.

At frame t, the input feature vector is $f_t$, and the hidden state of the GRU is denoted as $g_t$. The temporal attention score is obtained by two fully connected layers, as in formula (7).

$$a_t = \sigma\left(v^T \tanh(Mg_t + d) + D\right) \tag{7}$$

In formula (7), $a_t$ is the contribution of each frame in aggressive behavior recognition task, is used to control the output between [0, 1], $v \in R^{m\times 1}$ and D are the weight parameters and biases of the second-layer fully connected neural network, and $M \in R^{m\times n}$ and $d \in R^{m\times 1}$ are the weight parameters and biases of the first-layer fully connected neural network. In order to more clearly represent which frame is prominent or suppressed, the temporal attention score is normalized according to formula (8).

$$w_t = \frac{\exp(a_t)}{\sum_{t=1}^{T}\exp(a_t)} \tag{8}$$

where T is the number of input image frames, and $w_t$ is the temporal attention weight. Then according to formula (9), the temporal attention weight and the features of each frame are fused to generate the temporal attention feature F.

$$F = \sum_{t=1}^{T} w_t g_t \tag{9}$$

### 2.3.3. Aggressive behavior recognition

The temporal attention feature was converted into a 2-dimensional vector through fully connected layer. Then, the softmax function was used to convert all the elements of this 2-dimensional vector into the values within the interval (0, 1) and normalized these values (the sum of all values is 1). Finally, the output of the output layer is translated into action class probabilities, and the result of aggressive or non-aggressive behavior recognition is obtained.

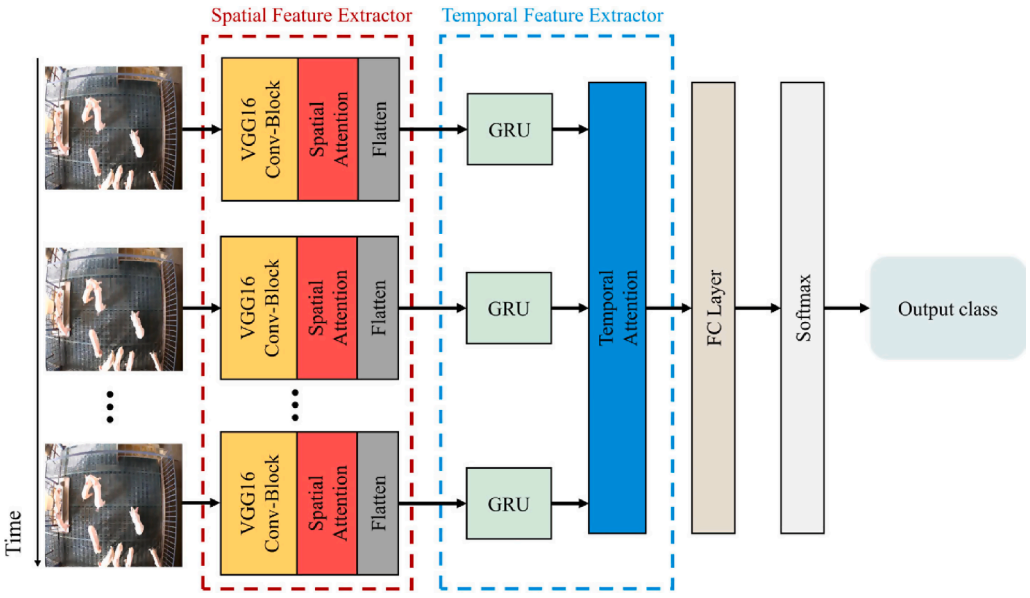


**Fig. 3.** Flowchart of the proposed hybrid model of aggressive behavior recognition.

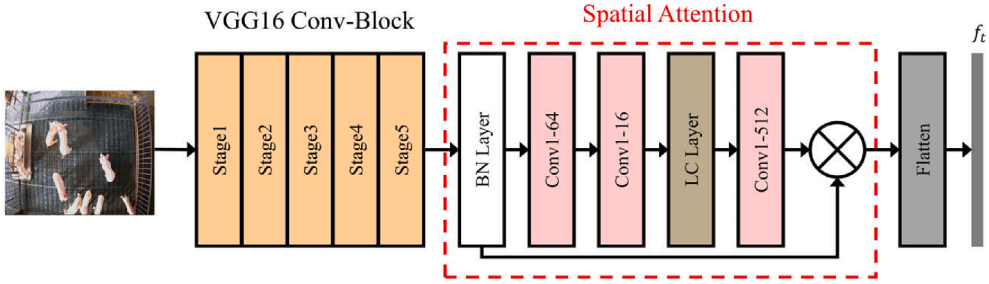


**Fig. 4.** Spatial feature extractor structure.

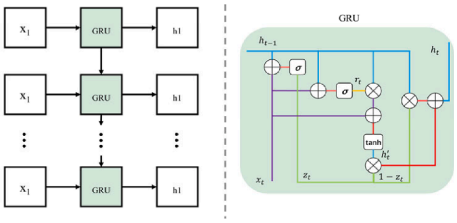


**Fig. 5.** Schematic diagram of GRU unit structure.

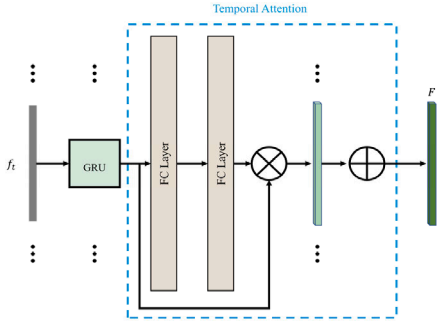


**Fig. 6.** Schematic diagram of temporal feature extractor structure.

### 2.4. Evaluation indicators

In this study, four evaluation indicators suitable for classification tasks, Recall, Precision, Accuracy and F1 Score, are used to measure the performance of the model. The formulas of the evaluation indicators are as follows:

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{10}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{11}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\% \tag{12}$$

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} \times 100\% \tag{13}$$

In the above formulas (10)-(13), TP represents the amount of data for which an aggressive behavior is correctly recognized as an aggressive behavior; TF represents the amount of data for which a non-aggressive behavior is correctly recognized as a non-aggressive behavior; FP represents the amount of data incorrectly recognized as aggressive behavior; FT represents the amount of data for which aggressive behavior was incorrectly recognized as non-aggressive behavior.

## 3. Results and analysis

### 3.1. Experimental settings

During the experiment, Python 3.6 was used to develop the algorithm, and all the algorithm models of this experiment were implemented on the Tensorflow 2.2.0 version framework. The experimental environment is shown in Table 3. The parameters of the aggressive behavior recognition model of group-housed pigs in this study are shown in Table 4.

### 3.2. Evaluation of functionality of each variant module

In order to prove that the spatial attention mechanism and the temporal attention mechanism can improve the performance of our CNN-GRU hybrid model in the spatial and temporal domains, respectively, the evaluation experiment is carried out in this study to prove the effectiveness of each module in the model. The following are the models of the experiment:

B1 (Baseline): A basic model for the recognition of aggressive behaviors of group-housed pigs based on video data. In this model, the VGG16 convolutional neural network with self-training model parameters is used as the spatial feature extractor to extract the spatial features of each frame of the input image, obtain the corresponding spatial feature vectors, and then use these spatial feature vectors as the input of GRU model to recognize the behavior of each video episode.

B2 (Baseline + SA): This variant model adds spatial attention mechanism (Spatial Attention, SA) to the spatial feature extractor part of B1.

B3 (Baseline + TA): This variant model adds temporal attention mechanism (Temporal Attention, TA) after the GRU model of B1.

B4 (Baseline + SA + TA): This variant model includes both SA and TA parts on the basis of B1.

The accuracy results of the B1-B4 models in the experiment on the test set are shown in Table 5. The batch size in this experiment is 32. It can be seen from the Table 5 that the recognition accuracy of the B1 model is the worst, and the recognition accuracy of the B2 model in the test dataset is 2.2 % higher than that of the B1 model, which proves that adding the spatial attention mechanism to the spatial feature extractor can optimize the spatial feature extraction. The performance of the B3 model is better than that of the B1 model, but the effect is worse than the B2 model, the B3 model only integrates the attention mechanism in temporal without optimizing the spatial feature extractor part, so the process of extracting spatial features cannot focus on the key area information in the spatial domain. Although the temporal feature extractor will focus on key frames with important features, the spatial features extracted in the B3 model will be doped with non-important information, resulting in its performance being relatively worse. From the comparison of the results in Table 5, it can be seen that the B4 model that integrates spatial attention and temporal attention mechanism proposed in this paper is better than other models in this comparative experiment, and the accuracy reaches 94.8 %, which can prove B4 model (our model) is valid.

**Table 3**
Experimental environment parameter table.

| Parameter settings | Version |
|---|---|
| Operating System | Windows 10 |
| CPU | Intel(R) Xeon(R) Gold 6244 |
| RAM | DDR3 16G*12 |
| GPU | NVIDIA RTX 5000 |
| Tensorflow | 2.2.0 |

**Table 4**
Parameters setting.

| Parameter name | Value |
|---|---|
| Initial learning rate | 0.00001 |
| Batch size | 32 |
| Iteration | 2000 |
| Optimizer | Adam |

**Table 5**
Performances of B1-B4 model on the test dataset.

| Model | Recall | Precision | Accuracy | F1 |
|---|---|---|---|---|
| B1 | 93.1 % | 89.4 % | 91.5 % | 91.1 % |
| B2 | 94.2 % | 93.1 % | 93.7 % | 93.6 % |
| B3 | 94.7 % | 91.3 % | 92.9 % | 92.9 % |
| B4 | 95.3 % | 94.2 % | 94.8 % | 94.7 % |

### 3.3. The effect of the position of the attention module

Theoretically, the spatial attention module can be placed after any convolutional block of a convolutional neural network model to capture key spatial information. According to existing research, the temporal attention module is usually placed after the recurrent neural network to capture important time-series information. Therefore, this paper does compare the performance of the spatial attention module in different positions.

A spatial attention module is added after the five convolution blocks of the VGG16 model for performance experiment, as shown in Fig. 7. The recognition performance exhibited by adding spatial attention modules after the first to fifth convolution blocks shows an overall upward trend. The model performance after adding the spatial attention module to the first to third convolutional blocks is lower than that of the model without spatial attention. The reason is that there is too much unimportant information in the low-level feature map, and the target behavior abstracted from the high-level feature map is more comprehensive. After the operation of the attention module, invalid attention weights will be generated, which will affect the ability of the model to extract key features. Finally, according to the performance test results, this paper adds a spatial attention module after the fifth convolution block of spatial feature extract.

### 3.4. Contribution evaluation of spatio-temporal attention module

This section will evaluate the contribution of the spatial and temporal attention modules in this paper. Fig. 8 shows the comparison of the heat map before and after adding spatial attention to the spatial feature extractor. It can be seen from the figure that the model with spatial

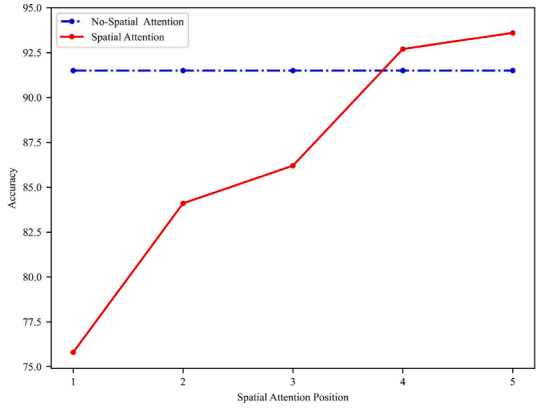


**Fig. 7.** Comparison the performance when placing the spatial attention module after different convolutional layers.

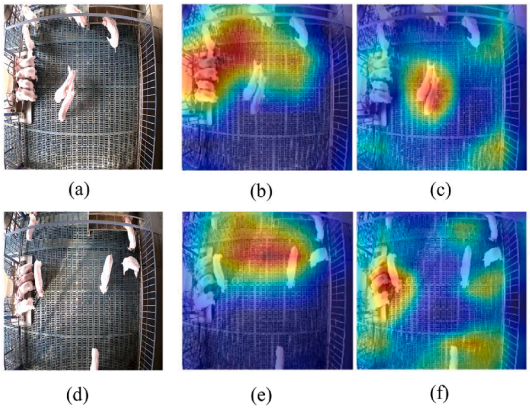


**Fig. 8.** The visualization of spatial attention weights: (a) is the original image of aggression, (b) is the weight heat map of (a) without spatial attention module, (c) is (a) with spatial attention module added, (d) is the original image of non-aggression, (e) is the weight heat map of (d) without spatial attention module, (f) is the weight heat map of (d) with spatial attention module.

attention is more concerned for the part of the image where aggressive behavior may occur and a relatively larger weight is assigned to this part, which will correspondingly suppress the influence of non-critical background information in the image on recognition of aggressive behavior. (b) and (e) are the models without the spatial attention module, and you can see that the features are mostly concentrated in some irrelevant locations. In c and f with the addition of the spatial attention module, the features are concentrated near the pig population. As shown by the results in Table 5, B2 (with the addition of the spatial attention module) has a certain performance improvement compared to B1 (without the addition of the spatial attention module).

In order to prove that temporal attention module in this paper has the function of automatically focusing on key frames, this paper visualizes the frame-by-frame weights of two aggressive behavior video episodes. Fig. 9 is a schematic diagram of the time series weights, in which the frame labels with obvious aggressive behavior characteristics are set to 1, and set the labels of other frames to 0. In Fig. 9, A and b represent two different video segments with a duration of 3 s. The red label represents whether the current frame is a significant fight behavior frame, the blue color represents the temporal attention weights extracted from the network, and the weights in the frame represent the temporal attention weights extracted from the whole model. The blue dots in the figure correspond to the weight of each frame respectively. It can be seen from the Fig. 9 that the weight corresponding to the frame labeled 1 is significantly higher than the weight corresponding to other frames. Therefore, the temporal attention used in this paper can assign higher weights to the key frames with the label 1 and achieve the purpose of automatically capturing important behavioral temporal features.

And, from Fig. 9, it shows that the area where aggressive behavior occurs is inconspicuous and the step what the behavior recognition contribution of each frame is different in time series, therefore it is very necessary to join the spatio-temporal attention module.

### 3.5. Comparison with other methods

For the field of behavior recognition, there are three most mainstream basic methods, namely 3D convolution method, CNN + LSTM method, and two streams method. Due to the extra data processing steps and huge memory consumption required for two streams networks, end-to-end training and applications are not possible. Therefore, most researchers choose the first two methods to apply to behavioral recognition in real environments.

Compared with the work of 3D CONVNet, this paper has a better recognition effect on the task of aggressive behavior of group-housed pigs. The 3D CONVNet network is improved by deepening the number of network layers and adding multi-scale feature fusion based on the 3D Convolutional Neural Network (C3D) network. Although it can be combined with high-level and low-level feature information, but there will still be some noise and invalid information in the pre-feature extraction network. In our model, an attention mechanism is added in both the spatial and temporal domains, which can effectively integrate spatiotemporal key information.

For the research of CNN + LSTM (Chen et al., 2020), the ImageNet pre-trained weight model is used as the spatial feature extractor, and the extracted spatial feature vectors are obtained through LSTM to obtain time series features to identify the aggressive behavior of pigs, ignoring the spatiotemporal saliency problem, and applied to the data of this paper on the set, its behavior recognition effect is limited. Further we use this method to recognize aggressive behavior on our dataset, and the results are shown in Table. 6. It can be seen that the accuracy of CNN + LSTM is lower than that the accuracy of 94.8 % achieved by this work.

### 3.6. Misclassification of the method in this paper

Fig. 10 shows an example of the misclassification of the method in this paper. In Fig. 10(a), the high-speed aggressive behavior is falsely classified as non-aggressive behavior. In video episode of this behavior, two pigs in the red box are fighting fiercely. However, this behavior occurs at the edge of the camera's field of view, so the key part of the

**Table 6**
Performances of different model on our test dataset.

| Model | Recall | Precision | Accuracy | F1 |
|---|---|---|---|---|
| 3D CONVNet | 89.6 % | 87.8 % | 88.7 % | 88.6 % |
| CNN-LSTM | 91.5 % | 93.0 % | 92.3 % | 92.2 % |
| Ours | 95.3 % | 94.2 % | 94.8 % | 94.7 % |

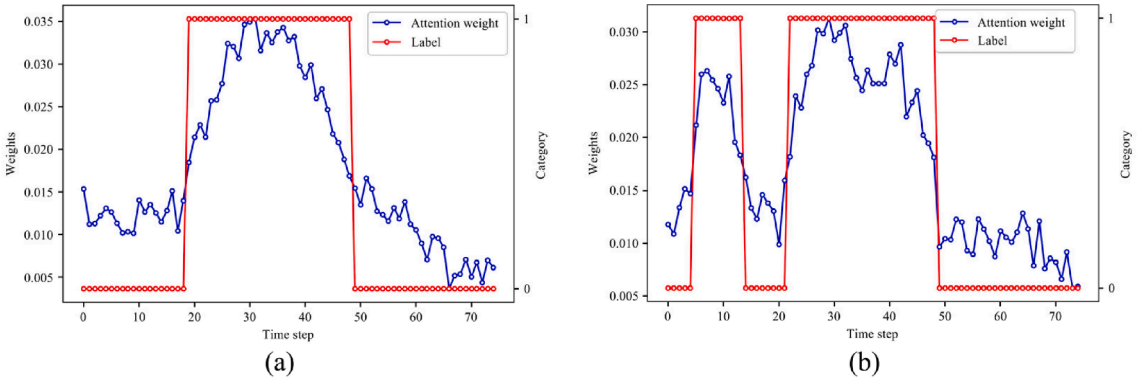


**Fig. 9.** The visualization of learned temporal attention weights.

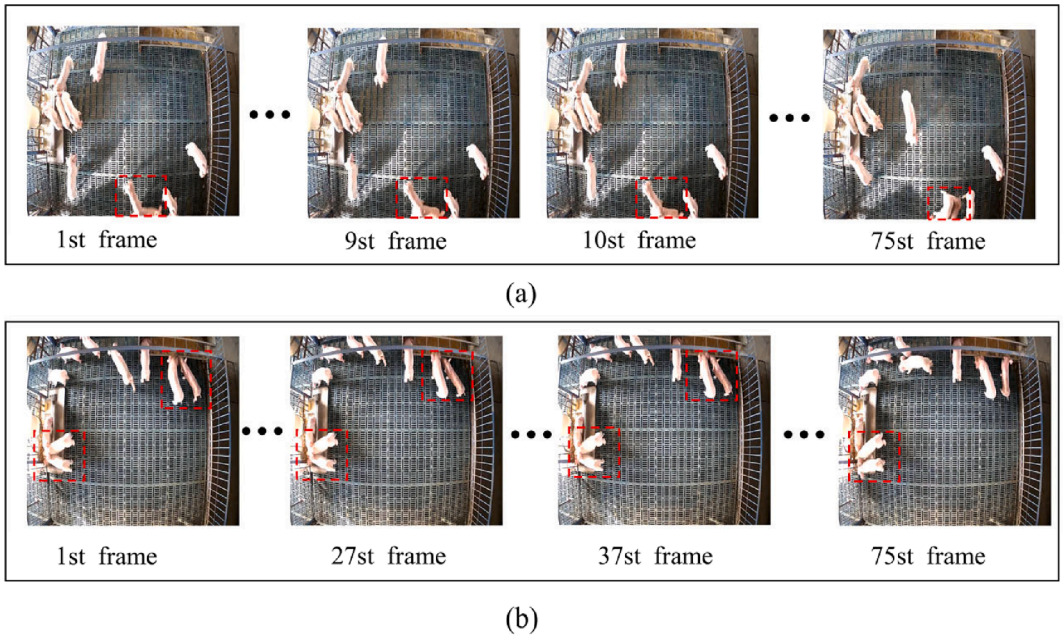


**Fig. 10.** Examples of misclassification:(a) is an example of aggression falsely recognized as non-aggression, (b) is an example of non-aggression falsely recognized as aggression.

aggressive behavior in the behavior period cannot be obtained, resulting in ambiguous behavior judgment, in addition, due to factors such as lighting and the angle of data collection, it cannot be accurately recognized. The mouth of the victim pig and the ears of the victim pig, or the part of the pig's body that is in contact with each other are continuously covered and have a similar tendency to avoid behavior as shown in Fig. 10(b). Under the influence of the above factors, behaviors with ambiguity are prone to errors in recognition. In the follow-up behavior recognition research, the strategy with the least external interference can be selected as far as possible for data collection.

## 4. Conclusion

In order to make full use of significant spatio-temporal information in the process of recognition the aggressive behavior of pigs, this paper designed a hybrid model consisted of the spatial feature extractor module with a spatial attention mechanism, and the temporal feature extractor with a temporal attention mechanism. Based on the dataset of this paper, the recognition accuracy of aggressive behavior of group-housed pigs is 94.8 %. In the comparative experiment of each variant module, the model that combines spatial and temporal attention modules in hybrid model of CNN and GRU has the best recognition performance, which proves the effectiveness of the model in this paper. In the experiments for evaluating the contribution of the spatial and temporal attention modules, it can be seen that the spatial attention module could guide the spatial feature extractor to assign relatively larger weights to local regions where occurred aggressive behavior, and when the temporal feature extractor obtained the temporal attention feature, relatively larger attention weights were given to the key frames of behavior recognition, and finally the effective feature for aggressive behavior recognition of group-housed pigs is obtained. Through the analysis of the misclassified examples of the method in this paper, we can consider in the future research to clearly capture the information such as the body parts of pigs in aggressive behavior with multi-camera data collection method, so as to improve the recognition effect of aggressive behavior of group-housed pigs.

## 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 github link to my data in the paper.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 31902210, 32172784 and 32072788, in part by the Natural Science Foundation of Heilongjiang Province of China under Grant QC2018074, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province of China under Grant UNPYSCT-2018142, in part by the Young Talents Project of Northeast Agricultural University under Grant 18QC23 and in part by China Agriculture Research System of MOF and MARA.

## References

Buettner, K., Scheffler, K., Czycholl, I., et al., 2015. Network characteristics and development of social structure of agonistic behaviour in pigs across three repeated rehousing and mixing events[J]. Appl. Anim. Behav. Sci. 168, 24–30.
Cho, K., Van Merriënboer, B., Bahdanau, D., et al. On the properties of neural machine translation: Encoder-decoder approaches[J]. arXiv preprint arXiv:1409.1259, 2014.
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., 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.
Dong, L., He, D., Chen, C., et al., 2021. Recognition of Aggressive Behaviour in Group-housed Pigs Based on ALR-GMM[J]. Nongye Jixie Xuebao/Tran. Chin. Soc. Agric. Machinery 52 (1), 201–208.
Dong, W., Zhang, Z., Song, C., et al., 2022. Identifying the Key Frames: An Attention-aware Sampling Method for Action Recognition[J]. Pattern Recogn. 108797.
Gao, Y., Chen, B., Liao, H.M., et al., 2019. Recognition method for aggressive behavior of group pigs based on deep learning[J]. Trans. CSAE 35, 192–200.

He, D.J., Liu, D., Zhao, K.X., 2016. Review of perceiving animal information and behavior in precision livestock farming[J]. Trans. Chin. Soc. Agric. Mach 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.

Li, Q.F., Li, J.W., Ma, W.H., et al., 2021. Research Progress of Intelligent Sensing Technology for Diagnosis of Livestock and Poultry Diseases[J]. Sci. Agric. Sin 54, 2445–2463.

Liu, D., Oczak, M., Maschat, K., et al., 2020. A computer vision-based method for spatial-temporal action recognition of tail-biting behaviour in group-housed pigs[J]. Biosyst. Eng. 195, 27–41.

Liu, L.S., Shen, M.X., Bo, G.Y., et al., 2014. Sows parturition detection method based on machine vision[J]. Nongye Jixie Xuebao= Transactions of the Chinese Society for Agricultural Machinery 45 (3), 237–242.

Neethirajan, S., 2017. Recent advances in wearable sensors for animal health management[J]. Sens. Bio-Sens. Res. 12, 15–29.

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.

Pei, W., Dibeklioğlu, H., Baltrušaitis, T., et al., 2019. Attended end-to-end architecture for age estimation from facial expression videos[J]. IEEE Trans. Image Process. 29, 1972–1984.

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

Song, W., Cai, W.Y., He, S.Q., Li, W.J., 2021. Dynamic graph convolution with spatial attention for point cloud classification and segmentation. J. Image Graph. 26 (11), 2691–2702.

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.

Sudhakaran, S., Escalera, S., Lsta, L.O., 2019:. Long short-term attention for egocentric action recognition[C]//Proceedings of the IEEE/CVF. Conference on Computer Vision and Pattern Recognition. 9954–9963.

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.

Yang, Q., Xiao, D., Cai, J., 2021. Pig mounting behaviour recognition based on video spatial–temporal features[J]. Biosyst. Eng. 206, 55–66.

Zhang, C., Chen, P., Lei, T., et al., 2022. What-Where-When Attention Network for video-based person re-identification[J]. Neurocomputing 468, 33–47.