Parallel channel and position attention-guided feature pyramid for pig face posture detection

Zhiwei Hu[1,2], Hongwen Yan[1*], Tiantian Lou[3]

(1. *College of Information Science and Engineer, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China*;
2. *School of Computer and Information Technology (School of Big Data), Shanxi University, Taiyuan 03006, China*;
3. *College of Agricultural Economics and Management, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China*)

Abstract: The area of the pig's face contains rich biological information, such as eyes, nose, and ear. The high-precision detection of pig face postures is crucial to the identification of pigs, and it can also provide fundamental archival information for the study of abnormal behavioral characteristics and regularities. In this study, a series of attention blocks were embedded in Feature Pyramid Network (FPN) for automatic detection of the pig face posture in group-breeding environments. Firstly, the Channel Attention Block (CAB) and Position Attention Block (PAB) were proposed to capture the channel dependencies and the pixel-level long-range relationships, respectively. Secondly, a variety of attention modules are proposed to effectively combine the two kinds of attention information, specifically including Parallel Channel Position (PCP), Cascade Position Channel (CPC), and Cascade Channel Position (CCP), which fuse the channel and position attention information in both parallel or cascade ways. Finally, the verification experiments on three task networks with two backbone networks were conducted for different attention blocks or modules. A total of 45 pigs in 8 pigpens were used as the research objects. Experimental results show that attention-based models perform better. Especially, with Faster Region Convolutional Neural Network (Faster R-CNN) as the task network and ResNet101 as the backbone network, after the introduction of the PCP module, the Average Precision (AP) indicators of the face poses of Downward with head-on face (D-O), Downward with lateral face (D-L), Level with head-on face (L-O), Level with lateral face (L-L), Upward with head-on face (U-O), and Upward with lateral face (U-L) achieve 91.55%, 90.36%, 90.10%, 90.05%, 85.96%, and 87.92%, respectively. Ablation experiments show that the PAB attention block is not as effective as the CAB attention block, and the parallel combination method is better than the cascade manner. Taking Faster R-CNN as the task network and ResNet101 as the backbone network, the heatmap visualization of different layers of FPN before and after adding PCP shows that, compared with the non-PCP module, the PCP module can more easily aggregate denser and richer contextual information, this, in turn, enhances long-range dependencies to improve feature representation. At the same time, the model based on PCP attention can effectively detect the pig face posture of different ages, different scenes, and different light intensities, which can help lay the foundation for subsequent individual identification and behavior analysis of pigs.

Keywords: objection detection, attention mechanism, feature pyramid network, face posture detection, pig
DOI: 10.25165/j.ijabe.20221506.7329

Citation: Hu Z W, Yan H W, Lou T T. Parallel channel and position attention-guided feature pyramid for pig face posture detection. Int J Agric & Biol Eng, 2022; 15(6): 222–234.

1 Introduction

The area of the pig face contains rich and valuable biological information that reflects the welfare status, well-being conditions, and social interactions[1,2]. Different facial postures contain information from different angles, such as the head-on face with pig nostrils, while only a single eye on the lateral face. Effective detection of the face area is conducive to the identification of pigs and can timely detect abnormal behavior according to different facial postures, further ensuring the health of pigs. Further significant impact on the economics of pig farms.

As modern farming continues to expand, continuous physical monitoring of animals is time-consuming, subjective, and impractical[3]. Some studies have turned to sensor techniques that can serve as an aid to the ability to automatically monitor biological responses in animals[4]. However, the sensing devices are invasive to the animal and can damage the animal's epidermis or internal organs. Computer vision-based techniques can provide non-contact, low-cost, and non-invasive methods, and have been widely used in livestock-related research[5].

In recent years, computer vision-based methods have attracted significant interest in the field of pig-related research, including, automatic recognition of pig posture[6], pig body composition estimation[7], pig aggressive behavior detection[8], piglet stress prediction[9], and pig target tracking[10]. However, when the textures of foreground and background objects are very similar, the above traditional methods are difficult to produce satisfactory results. The main reasons are the following two limitations: 1) Uneven light conditions, and environmental information such as urine stains, manure, pigpen, and other debris have brought great challenges to the research on pig targets. 2) The above methods mainly rely on morphological features such as color, shape, or texture for feature extraction, and the selection of thresholds or improper selection of feature equations will lead to the fluctuation

Received date: 2022-01-15 **Accepted date:** 2022-10-14

Biographies: Zhiwei Hu, PhD, Lecture, research interest: animal husbandry informationization, Email: zhiweihu@whu.edu.cn; **Tiantian Lou**, Master, Lecture, research interest: agricultural informatization, Email: sxaultt@163.com;
***Corresponding author: Hongwen Yan**, PhD, Associate Professor, research interest: agricultural informatization. Address: College of Information Science and Engineer, Shanxi Agricultural University, Jinzhong 030801, Shanxi, China. Tel: +86-16603543910, Email: yhwhxh@126.com.

of performance. Deep learning-based techniques can automatically extract information without the need to manually construct feature equations for key information extraction, and have achieved remarkable performance in many fields[11,12].

Convolutional neural network (CNN), as a representative deep learning technology, has strong feature extraction ability and has been widely applied to precision agriculture[13], such as the detection of dairy cows[14], goats detection[15], weeds identification[16], and other fields. In particular, CNN-based technology has also been applied in the fields of individual pigs, such as recognizing abnormal behaviors of pigs[17], classification of drinking and drinker-playing in pigs[18], segmentation of group-raised pigs[19-21], counting or tracking pigs[22], and detecting pigs standing and lying down postures[23,24]. However, the above-mentioned CNN-based studies on pig have two shortcomings: 1) From a research point of view, the above researches mainly focus on pig's overall posture or behavior detection, and there is no relevant literature on facial posture with richer biological information. 2) In terms of research methods, only the open-domain CNN models are directly or fine-tuned to the field of pigs, but there is a certain difference between pig images and other natural images in practice, which leads to limited performance improvement of the corresponding model.

Given the deficiencies in the fields mentioned above, the pig face area was selected with rich identification parts as the research subject. To the best of the general knowledge, there is currently no relevant literature on the detection of facial posture. The only related researches focus more on the recognition of pig facial identity information[1,2], but in practice, the acquisition of the pig face is inherently challenging. Comparing the two tasks, the detection of face posture can be regarded as forward-looking work for pig face recognition tasks. To adjust the structure of CNN, the attention mechanism is introduced into a variety of deep learning tasks. Tong et al.[25] used channel attention-based DenseNet[26] for scene classification of remote sensing images. Chen et al.[27] proposed a spatial attention residual network to repair the low-resolution facial structure. Besides, many researchers used the Convolutional Block Attention Module (CBAM)[28], Bottleneck Attention Module (BAM) [29], Spatial-Channel Squeeze & Excitation (SCSE)[30], and Dual Attention Network (DANet)[31] to recalibrate the feature map from the channel and spatial dimensions. Inspired by these works, the attention mechanism mainly based on three reasons was introduced in this study: 1) The attention-based mechanism can effectively distinguish the specific information, further focus on the regions that are conducive to the detection of face posture, and suppress the message as manure, and pigpen to improve the detection accuracy of the face; 2) The pig face area contains information about the nose, eyes, and ears compared with other parts. Paying more attention to these areas can improve the performance of face posture detection; 3) The attention mechanism has achieved good results in other open domains, so it is of practical significance to introduce it into the pig face posture.

A novel feature pyramid network (FPN) incorporating multiple attention modules was proposed to detect facial posture in group-housed pigs in this study. In particular, the Faster Region Convolutional Neural Network (Faster R-CNN), Cascade R-CNN, and Libra Faster R-CNN were chosen as task networks and ResNet50 and ResNet101 as backbone networks for detailed comparison experiments. In order to capture the channel dependencies and the pixel-level pairwise relationships, respectively, the channel and position attention blocks are further introduced into the FPN framework. Meanwhile, various ways of cross-fusion of the two attention blocks are deeply explored, and the Parallel Channel Position (PCP), Cascade Position Channel (CPC), and Cascade Channel Position (CCP) modules are separately constructed to fuse the channel and position attention information in a parallel or cascade manner, respectively. Finally, Faster R-CNN-R101 was taken as the experimental model to visualize the heatmaps at different stages of FPN before and after adding the PCP module, and visualize the prediction results at different ages, different degrees of adhesion, and different light intensities.

Overall, the contributions can be summarized as follows:

1) Three task networks and two backbone networks are selected for facial posture detection of group-housed pigs;

2) The channel and position-based attention blocks are designed to capture the channel dependency and the pixel-level pairwise relationship, respectively;

3) Explore the fusion of parallel and cascade manners to find the best combination of two kinds of attention information;

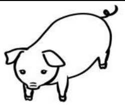
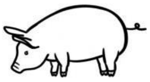
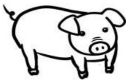
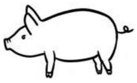
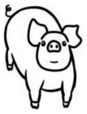
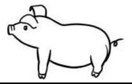
4) Visualize the attention heatmaps before and after adding the PCP module in FPN, and further visualize the prediction results of different stages of age and scenes to verify the robustness of attention mechanisms.

2 Definition of the pig face postures and data preparation

2.1 Definition of the pig face postures

The combination of the position, orientation, and connection relationship of various parts of the body is referred to as the posture, and if it is applied to the face area, the corresponding concept becomes the face pose. For individual pigs, the common facial postures are listed in Table 1. Among them, the "head position" is subdivided into "downward", "level" and "upward" according to the angle between the neck and the lowest point of the mouth, and "face position" is subdivided into "head-on" and "lateral" according to the number of pig eyes.

Table 1 The definition of the pig face postures

Posture category	Posture diagram	Posture description
Downward with head-on face (D-O)		Two eyes are in the field of view, and the angle is greater than −10°
Downward with lateral face (D-L)		Only one eye is in the field of view, and the angle is greater than −10°
Level with head-on face (L-O)		Two eyes are in the field of view, and the absolute angle is no more than 10 °
Level with lateral face (L-L)		Only one eye is in the field of view, and the absolute angle is no more than 10 °
Upward with head-on face (U-O)		Two eyes are in the field of view, and the angle is greater than +10 °
Upward with lateral face (U-L)		Only one eye is in the field of view, and the angle is greater than +10 °

Note: Different degrees represent the angle between the lowest point of the mouth and the horizontal direction of the neck.

2.2 Data preparation

2.2.1 Animals and housing

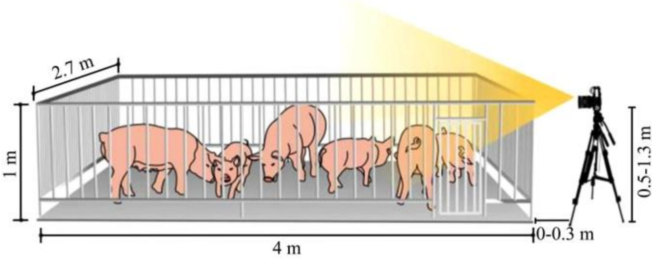
"The East Songjiazhuang Village Pig Farm" was selected

(Jicun Town, Fenyang City, Shanxi Province, China, it was defined as ESV-Farm) and the “the Experimental Animal Management Center of Shanxi Agricultural University” (Taigu City, Shanxi Province, China, it was defined as EAMC-Farm) as the experimental data collection base. The group-raised Edelschwein and Landrace mixed breed pigs were selected as research individuals and the body sizes of the pigs were different, the ages ranged from 20 to 105 d. Each pigpen of ESV-Farm measured approximately 8.75 m^3 (3.5 m×2.5 m×1.0 m) and for the EAMC-Farm, the area of pigpens was 10.8 m^3(4.0 m×2.7 m×1.0 m, as shown in Figure 1). In order to improve the generalization ability and robustness of models, the data were collected in two periods: June 1, 2019 (sunny with strong light, 23℃-29℃) and October 13, 2019 (cloudy with weak light, 10℃-19℃), respectively. A total of 45 pigs spread across eight pigpens were monitored from 9:00 to 14:00, and each pigpen was captured for a period of 64 min and 35 s videos. To obtain the horizontal view images, the camera (Canon 700D with anti-shake lens) was used and mounted on a tripod. The video frame rate was 25 fps and the resolution is 1920×1080 pixels. In order to ensure the continuity of the data, two videos with a duration of more than 30 min were selected as the initial data for each pigpen.

2.2.2 Data collection methods

Most studies used a top-view manner for pig data collection. Compared with the top-view acquisition method, the horizontal view was innovatively chosen to acquire experimental data. Taking EAMC-Farm as an example, the corresponding data collection scenario and its configuration parameters are shown in Figure 1. Group breeding pigs have the characteristics of poor movement trajectory controllability and strong adhesion. Therefore, the distance between the camera lens and the pigpens was adjusted to a range of 0 to 0.3 m. At the same time, in order to obtain images with different horizontal perspectives, the height of the camera tripod from the ground is set from 0.5 to 1.3 m.

Compared with the top-view data collection manner, the horizontal view method has the following advantages: 1) It can effectively capture detailed key parts, such as the face or hoof, which are more biologically informative. The top view mainly focuses on obtaining the back or abdomen information, which is inconsistent with the face posture detection task; 2) The distance between the lens and the pigs is convenient to adjust, and it is easy to obtain various condition data; 3) The horizontal viewing angle is consistent with the human observing angle of animals, so the model trained on the collected data is more suitable for transferring to mobile applications.

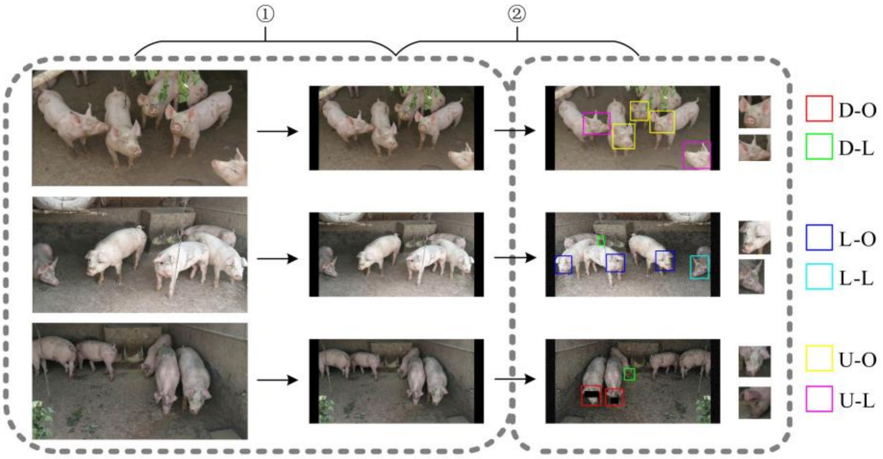


Note: The distance between the tripod and the pigpen is floating between 0-0.3 m, and the height of the tripod is floating between 0.5-1.3 m.

Figure 1 Platform of data collection for the test of this study

2.2.3 Data pre-processing

The collected videos cover different scenes, viewing angles, the intensity of lighting changes, and different age stages. To be able to obtain suitable model input, the following operations were performed on the acquired video to preprocess the datasets. The entire data process is shown in Figure 2.



Note: The first column represents the frame image cropped from the original video, the second column denotes the images after reducing the resolution size, and the third column means the labeling results after data enhancement processing.. The legend labels mean different postures of pigs. D-O: Downward with the head-on face; D-L: Downward with the lateral face; L-O: Level with the head-on face; L-L: Level with the lateral face; U-O: Upward with the head-on face; U-L: Upward with the lateral face.

Figure 2 Data pre-processing process of pig images of this study

1) Firstly, the videos were cropped every 25 frames to obtain 1933 images with a resolution of 1920×1080 pixels. Then, fix the aspect ratio to 2:1 to adjust the image size to 2048×1024 pixels, and fill the blank areas with black pixels. Afterward, in order to reduce the memory footprint, global scaling and coordinate transformation operations are performed to convert the resolution to 512×256 pixels. The whole process is shown in Figure 2.

2) In order to obtain more abundant scene data, data augmentation is performed on the basis of the limited amount of data obtained, thereby improving the robustness and generalization of the trained models. The data augmentation operations include four dynamic ways of brightening, mirroring, adding noise, or randomly occluding ways. For each augmented image, each augmentation operation is performed with a probability of 0.5 and contains one to four different transformations. For the brightening operation, randomly change the brightness value between 0.8 and 1.2, greater than 1 for dimming, and less than 1 for brightening. For mirroring, the horizontal mirror symmetric method was selected. For noise, the Gaussian noise was added. For random occlusion, the face coverage area was only randomized. The

whole process is shown in Figure 2.

3) After the above two steps, a total of 3866 images were obtained, which were randomly divided into training, validation, and test datasets, with corresponding numbers images of 1 933 579, and 1354 images respectively. Before entering the model, the mean value (123.675, 116.28, 103.53) and variance (58.395, 57.12, 57.375) of the three channels were taken to normalize the image values to speed up model convergence. In addition, the number of individuals with different face postures in the training, validation, and test datasets is shown in Table 2.

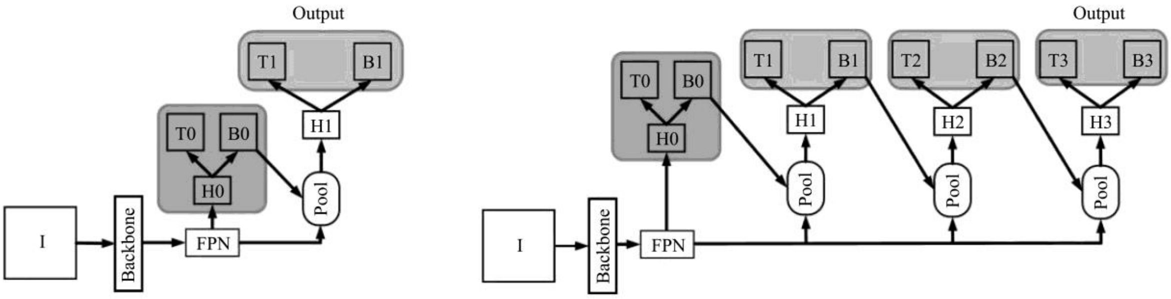
Table 2 Number of individuals of different facial postures

Datasets	D-O	D-L	L-O	L-L	U-O	U-L
Train	722	1512	950	2027	494	747
Val	202	463	311	586	160	269
Test	465	1086	727	1442	379	535

2.3 Pig face posture detection model

2.3.1 Three types of task models

Given the success of the R-CNN architecture, the combination of proposal-based detectors and region classifiers has dominated the two-stage approach to object detection problems[32,33]. The Faster R-CNN[34] further accelerates by introducing a Region Proposal Network (RPN)[34], and its simplified model structure architecture is shown in Figure 3a. To alleviate the contradiction of scale mismatch between the RPN receptive fields and actual object size, the FPN was introduced[35] into the Faster R-CNN to detect proposals at multiple scales. In general, the performance of Faster R-CNN is highly dependent on the selection of the IOU (Intersection Over Union)[36] threshold, setting a larger IOU threshold will lead to an exponential decrease in the number of positive samples, and setting a smaller IOU threshold will bring more noise sample, the selection of the threshold is a hyperparameter, which is difficult to obtain in practice. Cascade R-CNN[37] can be used to address this problem, as shown in Figure 3b, in the sequential training process in different stages, the output of the previous stage is used as the training input of the next stage. The output of a detector trained with a smaller IOU threshold favors the training of a detector with the next higher IOU threshold. Cascade R-CNN consists of four stages, one RPN and three detectors with different IOU thresholds {0.5, 0.6, 0.7}, corresponding to H1-H3 in Figure 3b. However, the above Faster R-CNN and Cascade R-CNN cannot solve the problem of sample imbalance. The Libra R-CNN[38] proposes IOU-balanced sampling, which mines hard samples according to IOU with assigned ground truth to address the sample-level imbalance problem, and its model architecture is similar to Faster R-CNN.



a. Faster R-CNN

b. Cascade R-CNN

Note: I, FPN, Pool, H*i*, B*i*, RPN, and T*i* (*i*=0, 1, 2, 3) denote input image, feature pyramid network, region-wise feature extraction, network head, bounding box, region proposal network, classification, respectively. “B0” is the proposal in all architecture.

Figure 3 Architecture of Faster R-CNN and Cascade R-CNN

2.3.2 Feature pyramid network

Feature context information is quite important for the detection of pig face posture. Existing CNN models usually learn deep features of the object by stacking multiple convolutional and pooling layers. However, the area of the pig face has large variations in scale, shape, and location. Traditional methods usually directly use bottom-up convolution and pooling operations, which may be difficult to effectively deal with the challenges posed by complex changes in face regions. For the feature extracted by the stacked convolutional network, the low-level features lack semantics but have rich location contents, while the high-level features have rich semantic information but lack location knowledge. Fusion of high and low-level features can effectively alleviate the contradiction between semantic and location information that cannot be improved at the same time. FPN (a component in Figure 3) can extract feature maps of different scales at multiple levels, and its construction involves a bottom-up pathway, a top-down pathway, and lateral connections. The FPN architecture combines low-resolution, semantically strong features with high-resolution, semantically weak features through the top-down pathway and lateral connections. The FPN model structure using ResNets[39] as the backbone network is shown in Figure 4.

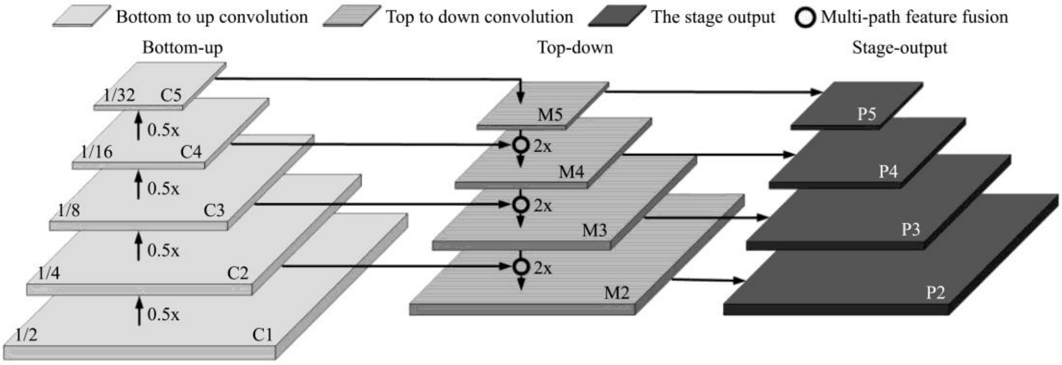
2.3.3 Feature fusion with different attention modules

A novel pig face posture detection method was proposed which consists of a pyramid feature extraction module (as shown in Figure 4) and a series of attention modules to capture context-aware multi-scale multi-receptive field features. The component architecture with different attention modules is shown in Figure 5 (corresponding to the black and bold open circles in Figure 4).

Local feature extractor. The standard convolutional layer learns the local feature from the eight adjacent feature vectors, corresponding to the red dashed area in Figure 5a, it was denoted as $f_{\mathrm{loc}}(*)$ in this study. To obtain richer semantic information, two convolutional layers were performed with corresponding convolution kernel sizes of 1×1 and 3×3, respectively. In Figure 5b, for the red dot region, the standard convolution can obtain the semantic information related to the pig’s eyes. Therefore, local features are beneficial to the extraction of deep semantic features.

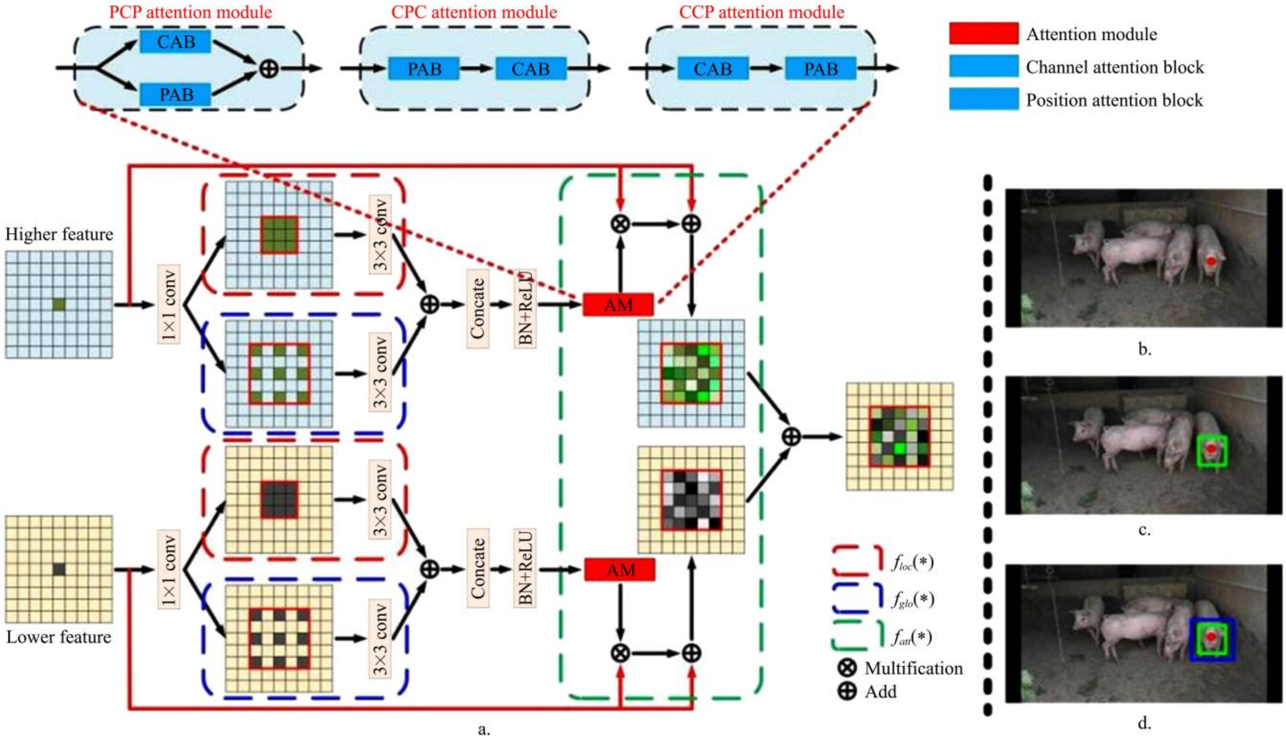
Global context extractor. Different receptive field sizes are helpful for information extraction for objects of different sizes. In order to increase the receptive field and reduce the amount of calculation, the conventional convolution operation adopts the down-sampling operation, however, it will lead to the problem that the spatial resolution is significantly reduced. Atrous dilated convolution has relatively large receptive fields and can effectively learn the surrounding context without losing the resolution and adding extra parameters. Inspired by Context Guided Network

(CGNet)[40], a 3×3 atrous convolutional layer was adopted as the global context extractor $f_{\mathrm{glo}}(*)$, corresponding to the blue dashed area in Figure 5a. As shown in Figure 5b, when the red areas were focused on, it is difficult to classify the categories to which the red dots belong. While in Figure 5c, the green box region is the surrounding context of the red dot, it is easier to identify the green area of the pig face when both the red dot and its surrounding context are considered. For Figure 5d, if the receptive field (blue region) is expanded, the more global context as well as the red dot and its surrounding (green region) can be further captured. In addition, the receptive fields of different scales can better separate the adhering pig face regions.



Note: 0.5x represents down-sampling operation; 2x represents up-sampling operation; C*i*, M*i*, and P*i* represent the corresponding the feature map of the *i*th component block of Bottom-up, Top-down, and Stage-output, respectively.

Figure 4 Architecture of feature pyramid network



a. Attention component b-d. An example to prove the importance of global information

Note: Conv: Convolution; BN: BatchNorm; ReLU: Rectified Linear Unit; $f_{\mathrm{loc}}(*)$: Local feature extractor; $f_{\mathrm{glo}}(*)$: Global context extractor; $f_{\mathrm{att}}(*)$: Attention module.

Figure 5 An overview of the attention component with different attention modules

Attention module. It is significant to perform deep filtering on features extracted by local or global context extractors. The standard convolutional network simply and linearly superimposes features from different sources, which usually leads to redundant information and degrades prediction performance. The attention mechanisms enable the neural network to focus more on the regions that are beneficial to the task and exert better attention on the key regions, where channel and position attention can capture the channel dependencies and the pixel-level pairwise relationship, respectively. To effectively combine the two types of attention mechanisms, a series of attention modules were implemented, such as the Parallel Channel Position (PCP) attention module, Cascade Position Channel (CPC) attention module, and Cascade Channel Position (CCP) attention module. PCP fuses the two-way parallel attention information, while CPC and CCP are cascade attention patterns to explore the degree of influence of channel and position attention execution order on detection results.

2.3.4 Channel attention block (CAB)

Each channel of a feature map can be considered as a response to a specific class, different channels respond to different semantic categories, and the responses between different semantic responses are associated with each other[41]. The channel attention mechanism can emphasize the interdependence between feature maps and improve the feature representation for specific classes by exploiting the inter-interdependencies among channels. Inspired by DANet[31] and CBAM[28], An enhanced channel attention block was constructed to explicitly encode inter-channel dependencies. The channel attention block consists of two stages (Stage #1 and Stage #2), and the specific operations are shown in Figure 6.

Stage #1 can be summarized in two procedures.

1) Given a local feature map $\boldsymbol{I}\in\boldsymbol{\Omega}^{C\times H\times W}$ first reshapes $\boldsymbol{I}$ to $\boldsymbol{\Omega}^{C\times HW}$ and performs a convolution operation with a kernel size of 1×1 to generate the feature map $\boldsymbol{A}\in\boldsymbol{\Omega}^{l\times H\times W}$, where C, H, W represent the number of channels and the height and width of the feature map, respectively. Subsequently, $\boldsymbol{A}$ is reshaped and transposed to $\boldsymbol{B}\in\boldsymbol{\Omega}^{HW\times1\times1}$. Next, a matrix multiplication was performed between reshaped $\boldsymbol{I}$ and $\boldsymbol{B}$. Finally, apply the softmax function to obtain the channel-wise attention weights. The operations are defined as follows.

$$x_t=\frac{exp(W_t)}{\sum_{s=1}^{C}exp(W_s\cdot W_t)} \tag{1}$$

$$W_t=\sum_{i=1}^{HW}I_{t,i}\cdot B_i \tag{2}$$

where, $I_{t,I}$ is the tth row and ith column of the reshape $\boldsymbol{I}$, W_t measures the tth channel's value impacted by the reshaped $\boldsymbol{I}$, and $\boldsymbol{B}$. x_t is the tth channel's value after softmax operation.

2) Perform an element-wise multiplication between the input feature $\boldsymbol{I}$ and the channel attention map $\boldsymbol{X}$ to get the result $F_{stg\#1}^{cab}\in\Omega^{C\times H\times W}$.

Stage #2 can be abstracted into two procedures as follows:

1) First, the input feature $\mathbf{I}$ was reshaped to $\boldsymbol{\Omega}^{C\times HW}$, and then perform a matrix multiplication between $\mathbf{I}$ and the transpose feature $\mathbf{I}$. Then, apply a softmax layer to obtain the channel attention map $\mathbf{Y}\in\boldsymbol{\Omega}^{C\times C}$. The element $y_{t,s}$ of $\mathbf{Y}$ denotes the sth channel's impact on the tth channel.

$$y_{t,s}=\frac{exp(I_s'\cdot I_t)}{\sum_{s=1}^{C}exp(I_s'\cdot I_t)} \tag{3}$$

The denominator can map the interdependent values from 0 to 1.

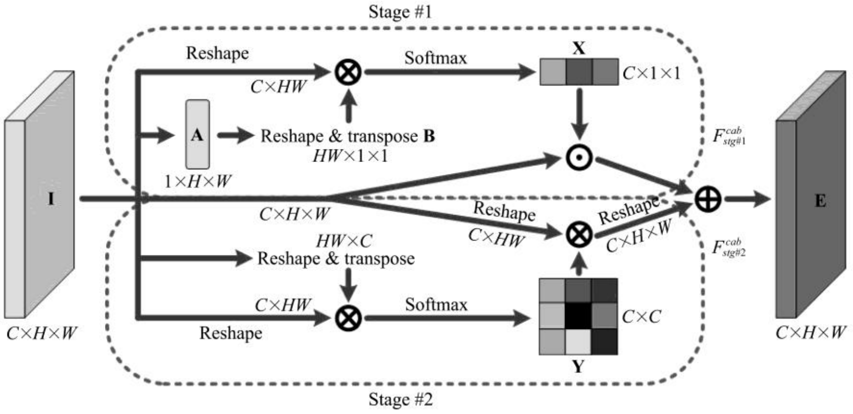
2) Next we perform a matrix multiplication between the transpose of $\mathbf{Y}$ and the reshape $\mathbf{I}$, further reshaping the result $F_{stg\#2}^{cab}\in\Omega^{C\times H\times W}$.

The element-wise addition was applied between Stage #1 and Stage #2 output, further getting the results of CAB as follows.

$$F_{cab}=F_{stg\#1}^{cab}+F_{stg\#2}^{cab} \tag{4}$$

2.3.5 Position attention block (PAB)

Different from channel attention, position attention focuses on “where” information, which is more refined than channel attention and can complement channel attention. In order to model rich contextual content over local features, inspired by DANet[31] and CBAM[28], an enhanced position attention block (PAB) was introduced. The PAB is able to encode wider contextual information into local features, further enhancing the feature expression ability. The position attention block consists of two stages (Stage #1 and Stage #2) as in Figure 7.



Note: The feature maps are denoted as feature dimensions, *e.g.* C×H×W denotes a feature map with channel number *C*, height *H*, and width *W*. ⊗ represents matrix multiplication, ⊙ denotes element-wise multiplication, and ⊕ denotes element-wise addition.

Figure 6 An overview of the channel attention block

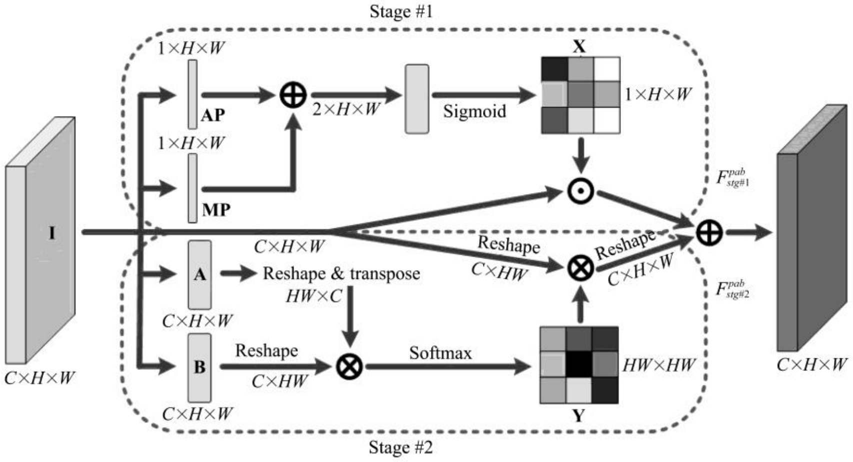


Figure 7 An overview of the position attention block. The symbol annotations are consistent with Figure 6

To compute the position attention of Stage #1, the following two steps were applied:

1) the average-pooling AP and the max-pooling MP were the first performed operations along the channel dimension and

concatenate the corresponding results together to generate efficient feature fusion results. Then, apply a convolutional layer operation with a convolution kernel size of 3×3 to generate the position attention map $\boldsymbol{X}\in\boldsymbol{\Omega}^{l\times H\times W}$. The entire process of position attention calculation is shown in Equation (5).

$$F_x = \sigma(conv_{3\times3}[F_{ap};F_{mp}]) \tag{5}$$

where, F_{ap} and F_{mp} denote the average-pooling and max-pooling, respectively. $\mathrm{conv}_{3\times3}$ represents the convolutional operation with the kernel size 3×3. σ represents the sigmoid activation function. F_x denotes the attention weights of **X**.

2) Then, an element-wise multiplication was perfomed between the input feature **I** and the position attention map **X** to get the result $F_{stg\#1}^{pab} \in \Omega^{C\times H\times W}$.

Stage #2 can be summarized into two procedures:

1) Given the local feature map $\mathbf{I}\in\boldsymbol{\Omega}^{C\times H\times W}$, Stage #2 first applies two convolutional layers with the kernel size of 3×3 on **I** to generate two independent feature maps **A** and **B**, respectively, where {**A**, **B**}$\in\boldsymbol{\Omega}^{C\times H\times W}$. Then **A** and **B** were reshaped to {**A**, **B**}$\in\boldsymbol{\Omega}^{C\times HW}$. After that, a matrix multiplication was performed between the transpose of **A** and the feature map **B**, and further apply a softmax activation function to get the position attention map $\mathbf{Y}\in\boldsymbol{\Omega}^{HW\times HW}$, the acquisition of each attention value in the attention map **Y** is shown in Equation (6):

$$y_{t,s} = \frac{exp(A_s' \cdot B_t)}{\sum_{s=1}^{HW} exp(A_s' \cdot B_t)} \tag{6}$$

where, $y_{t,s}$ represents the sth position impact on the tth position.

2) A matrix was performed multiplication between the reshape of **I** and the position attention map **Y**, then reshape the result $F_{stg\#2}^{pab} \in \Omega^{C\times H\times W}$.

The element-wise addition was further conducted in Stages #1 and #2 to get the real output of the PAB block.

$$F_{pab} = F_{stg\#1}^{pab} + F_{stg\#2}^{pab} \tag{7}$$

3 Experiment

3.1 Implementation details and evaluation metrics

3.1.1 Implementation details

All experiments are implemented on PyTorch and mmdetection[42] for fairness of comparison The backbone networks used in the experiments are ResNet50 and ResNet101, which are abbreviated as R50 and R101 for clarity. For the train detectors with 16 GB Tesla V100 GPU, the batch size was set to 8, a total of 12 epochs were trained, and the initial learning rate was set to 0.02. All models are trained by Stochastic Gradient Descent (SGD) optimizer with a weight decay parameter of $1e^{-4}$ and momentum set to 0.9.

3.1.2 Evaluation metrics

The standard VOC-style[43] AP was selected as the evaluation metric. In order to measure the overall performance, the six types of face posture AP were averaged to get mean AP (mAP). The definitions of AP and mAP are as follows:

$$\mathrm{P} = \frac{tp}{tp + fp} \tag{8}$$

$$\mathrm{R} = \frac{tp}{tp + fn} \tag{9}$$

$$\mathrm{AP} = \int_0^1 \mathrm{P}\cdot\mathrm{R}dr \tag{10}$$

$$\mathrm{mAP} = \frac{1}{C}\sum_{C_i \in C} \mathrm{AP}_{(C_i)} \tag{11}$$

where, tp represents the number of detection boxes for which both the prediction and ground truth are positive samples. fp represents the number of detection boxes that are predicted to be positive samples but are actually negative samples. fn represents the number of detection boxes that predicted results are negative samples but actually positive samples. Taking D-O as an example, categories D-L, L-O, L-L, U-O, and U-L represent negative samples, and category D-O represents positive samples. P and R represent the precision and the recall of specified categories, respectively. C denotes the total number of pig face postures, C was set to six here.

3.2 Detection performance of different backbone networks and task networks with different attention blocks

3.2.1 Compare with different attention blocks in AP index

Different backbone networks can extract feature information of different depths. To explore the influence of backbone networks on various task networks, two basic backbones were selected, ResNet50 and ResNet101[39], and three task networks, Faster R-CNN, Cascade R-CNN, and Libra Faster R-CNN to conduct experiments. In addition, the effects of adding CBAM[28], BAM[29], and DANet[31] attention modules were shown to FPN. Furthermore, the results of different attention blocks were presented to explore their effectiveness, and the corresponding results are listed in Table 3.

Different Backbone Networks: Although deeper pre-trained network models based on ResNet101 may be slightly inferior to models based on ResNet50 in some pose classifications, overall the ResNet101-based model with or without attention blocks outperformed models based on ResNet50 with the same task networks. Taking Faster R-CNN as the task network as an example, using ResNet101 produces a result of 83.73% in mAP, which brings a 2.77% improvement over using the ResNet50 backbone network. For different pig face posture categories, compared with ResNet50-based Faster R-CNN, the ResNet101-based Faster R-CNN significantly improves the performance by 2.22%, 3.6%, 5.76%, 5.32%, 7.97%, and 5.21% in D-O, D-L, L-O, L-L, U-O, and U-L, respectively. The main reason is that ResNet101 contains more layers, which can extract stronger semantic and richer contextual information.

Different Task Networks: Compared with Faster R-CNN, using Cascade R-CNN as the task network and using ResNet50 and ResNet101 as the backbone network can stably bring 1.5% and 0.99% mAP improvement, respectively. Specific to the single pig face posture category, the performance of Cascade R-CNN has been improved more than that of Faster R-CNN. Specifically, on the D-O, D-L, L-O, L-L, U-O, and U-L categories, the improvements are 1.55%, 1.45%, 0.09%, 1.15%, 2.35%, and 2.43%, respectively. The reason is that, compared to Faster R-CNN, the Cascade R-CNN consists of a series of detectors that are trained in series by increasing the IOU thresholds, using the output of the previous detector as the input of the next higher quality detector, and continuously filter out negative samples to improve performance.

Compare with existing attention models: Adding CBAM, BAM, and DANet attention blocks in existing research to FPN can improve the model performance to a certain extent. Compared with CBAM and BAM, adding DANet can bring greater improvement. Specifically, using Faster R-CNN as the task network and ResNet50 as the backbone network, in terms of mAP, adding DANet is 1.69% and 0.76% higher than adding CBAM and BAM, respectively. In addition, the effect of the BAM attention block is better than the CBAM attention block. This is mainly because BAM uses a parallel manner to combine spatial and channel attention, while CBAM uses a cascade combination manner, which is prone to cascade errors, causing the accumulation of error information to affect the experimental effect. Although CBAM, BAM, and DANet attention modules can all improve the original model, the proposed PCP module in this study can surpass the above three attention modules under various backbone network and task network combinations, proving the effectiveness and robustness of the PCP module.

Table 3 Different task networks' performance under different attention blocks

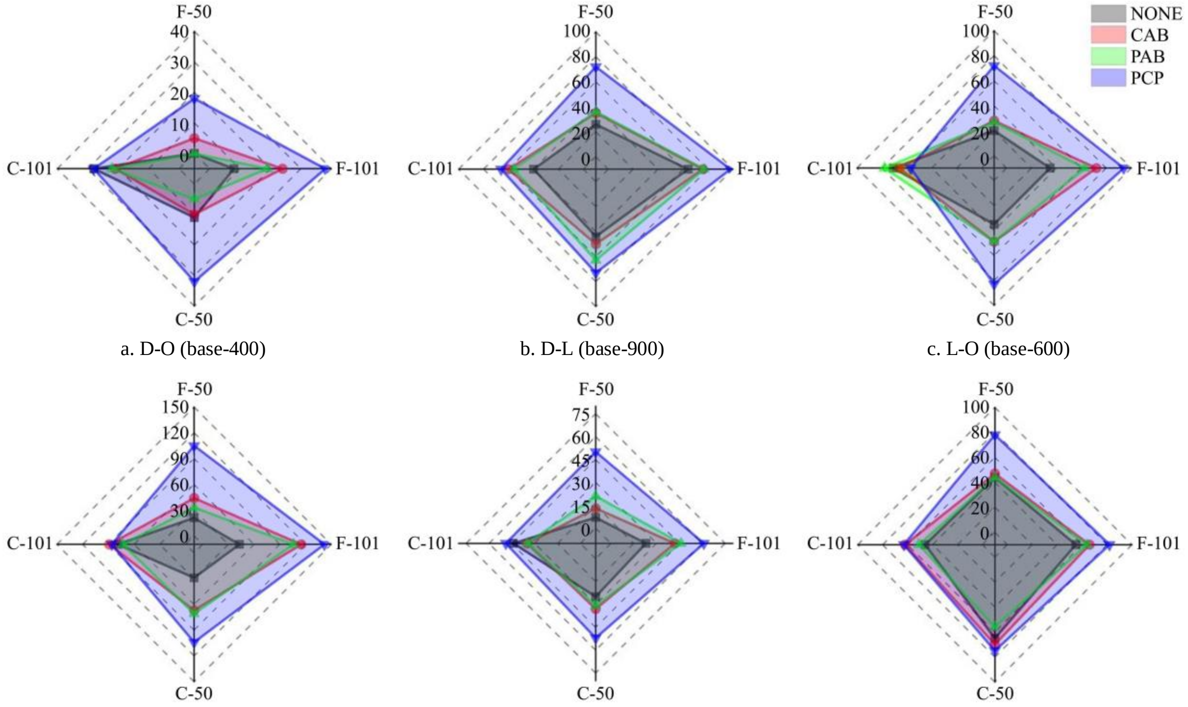
Model	Backbone	Attention	D-O	D-L	L-O	L-L	U-O	U-L	mAP
Faster R-CNN	R50	NONE	83.83	83.93	80.50	81.35	75.06	81.07	80.96
		CBAM	84.36	84.28	82.30	83.24	77.49	82.37	82.34
		SCSE	84.98	85.12	83.14	83.67	80.18	82.55	83.27
		DANet	85.33	85.18	83.96	85.02	81.24	83.47	84.03
		CAB	84.91	84.66	81.73	82.96	74.82	81.35	81.74
		PAB	83.71	84.37	80.49	81.99	77.33	80.78	81.45
		PCP	**86.60**	**86.98**	**86.26**	**86.67**	**83.03**	**86.28**	**85.97**
	R101	NONE	86.05	87.53	84.01	83.41	77.87	83.52	83.73
		CBAM	89.32	88.60	88.22	88.95	83.28	85.29	87.28
		BAM	89.88	89.23	88.79	88.92	84.17	85.92	87.82
		DANet	90.34	89.46	89.67	89.38	84.76	86.32	88.32
		CAB	89.37	88.49	87.76	88.60	82.29	85.13	86.94
		PAB	88.00	88.49	87.01	87.38	83.65	84.98	86.59
		PCP	**91.55**	**90.36**	**90.10**	**90.05**	**85.96**	**87.92**	**89.32**
Cascade R-CNN	R50	NONE	85.38	85.38	80.59	82.50	77.41	83.50	82.46
		CBAM	86.38	86.43	84.30	85.62	80.23	85.62	84.76
		BAM	86.42	86.92	84.22	85.99	80.66	85.48	84.95
		DANet	87.66	87.21	85.68	86.18	82.13	85.79	85.78
		CAB	86.20	85.65	83.75	84.66	79.78	85.52	84.26
		PAB	85.61	86.73	83.70	85.18	78.81	83.10	83.86
		PCP	**89.67**	**87.57**	**87.57**	**86.40**	**84.56**	**86.11**	**86.98**
	R101	NONE	88.89	84.30	87.28	84.86	81.57	81.40	84.72
		CBAM	89.12	87.00	87.66	85.93	81.42	82.09	85.50
		BAM	89.24	87.23	87.78	86.14	81.59	82.48	85.74
		DANet	89.66	87.18	88.03	86.78	82.31	83.11	86.18
		CAB	88.94	87.33	86.46	**87.09**	81.20	84.77	85.97
		PAB	89.08	86.96	**88.36**	85.92	81.27	82.32	85.65
		PCP	**90.21**	**87.58**	86.10	86.57	**83.84**	**84.83**	**86.52**
Libra Faster R-CNN	R50	NONE	88.49	**89.05**	76.72	81.51	75.37	83.16	82.38
		CBAM	88.22	86.92	80.21	84.33	77.45	82.52	83.28
		BAM	88.37	86.93	80.85	84.68	77.62	82.33	83.46
		DANet	88.46	87.01	81.33	84.78	78.54	82.86	83.83
		CAB	87.77	85.71	**83.99**	84.27	**81.62**	82.62	84.33
		PAB	**88.57**	87.06	83.17	**85.65**	78.59	81.52	84.09
		PCP	88.18	87.25	83.62	85.58	79.07	**83.94**	**84.61**
	R101	NONE	85.96	86.71	85.25	86.40	80.92	**85.87**	85.18
		CBAM	87.56	88.38	86.22	85.86	81.04	83.96	85.50
		BAM	87.81	88.23	86.59	85.92	81.28	84.14	85.66
		DANet	88.42	88.46	86.92	86.11	82.16	84.39	86.08
		CAB	89.08	88.37	86.01	86.62	**83.05**	83.81	86.16
		PAB	**89.94**	**88.58**	86.39	85.48	80.40	84.86	85.94
		PCP	89.05	88.10	**87.55**	**86.71**	81.80	84.93	**86.36**

Note: R50 represents ResNet50, R101 represents ResNet101, and bold indicates the corresponding optimal value. NONE means do not add attention information, CBAM denotes Convolutional Block Attention Module, SCSE denotes Spatial-Channel Squeeze & Excitation, DANet represents Dual Attention Network, CAB means Channel Attention Block, PAB means Position Attention Block, PCP means Parallel Channel Position, and mAP denotes mean Average Precision.

Effectiveness of Different Attention Blocks: To explore the effect of proposed attention blocks, the performance of adding different attention blocks into FPN for pig face posture detection was further investigated. From Table 3, it can be seen that the attention-based models all outperform the models without attention blocks, and the CAB block performs slightly better than PAB blocks\ in the mAP metric. Using Cascade R-CNN as the task network, when using ResNet50 and ResNet101 as backbone networks, the model with CAB block improves the mAP metric over the PAB by 0.4% and 0.32%, respectively. For L-L and U-L face pose categories, with Cascade R-CNN-R101 as the baseline, compared to the PAB, adding the CAB block improves the AP metric by 1.17% and 2.45%, respectively. The main reason is that the CAB block assigns more weights to the channels that favor the facial posture category. Channel attention pays more attention to the "what" of the given input image, and at the same time can encode the correlation of channels, which further helps to improve the prediction accuracy of the pose category corresponding to the detection box. Different from channel attention, position attention is more focused on the acquisition of "were" information, which can improve the fineness of bounding box location detection. The reason why CAB is better than PAB is that CAB focuses on the category corresponding to the detection box, while PAB pays more attention to the position accuracy of the detection box. The boost in position only makes sense if the detection box category is correct. The performance can be further improved by integrating CAB and PAB in parallel. This demonstrates that the PAB block and CAB block can achieve complementary effects while capturing long-range dependencies in channel and position dimensions, thereby improving location detection accuracy while improving detection box categories.

3.2.2 Compare with different attention blocks in true positive values

It can be seen from Section 3.1.2 that the evaluation index AP is affected by the true positive value. ResNet50 and ResNet101 were used as backbone networks to count the true positive values on the Faster R-CNN and Cascade R-CNN task networks with different attention blocks, corresponding to the results shown in Figure 8. In order to obtain the true positive and false positive values in Equation (8), a total of three steps are performed: 1) First, filter the detection boxes at a certain position, remove the boxes whose confidence is less than a specified threshold, and remove the boxes whose predicted category is inconsistent with the real category; 2) Then sort the filtered prediction boxes in descending order according to the confidence score, and calculate the IOU value of the box with the highest confidence and the real box. If the IOU value is greater than the specified threshold, true positive values are added, and the mark of the real box is detected. 3) Add the remaining prediction boxes to the list of false positives.



a. D-O (base-400)

b. D-L (base-900)

c. L-O (base-600)

d. L-L (base-1200)

e. U-O (base-290)

f. U-L (base-400)

Note: The subgraph title base-i means that i is the base true positive value, and the true value of the corresponding subgraph true positive needs to be plus i. F-50 represents Faster R-CNN with ResNet50, and C-50 represents the Cascade R-CNN with ResNet50 backbone network.

Figure 8 TP values predicted by Faster R-CNN and Cascade R-CNN with different attention blocks for six facial postures

True Positive Values of Different Attention Blocks: For the coverage area of the radar map, the largest area (shaded blue) can be covered with the PCP block. Under the same conditions, the true positive values based on the attention block significantly outperform the model without adding attention information. As for the L-L face posture category, when selecting Faster R-CNN as the task network, and ResNet101 as the backbone network, it can be observed that attention-based models are significantly better than the model without adding attention blocks, and the true positive values after adding CAB, PAB and PCP attention reach 115,106, and 142, respectively, compared with no attention added, the increased is 72, 63, and 99, respectively. For different attention blocks, CAB blocks slightly outperform PAB blocks. Taking the L-O face posture as an example, compared with the baseline model Faster R-CNN with ResNet50 backbone, the true positive value of adding the CAB block is 28, which is slightly higher than after adding PAB attention. When the channel and position blocks are paralleled together, the true positive values achieve 72, which further significantly improves the performance. Therefore, it is important to fuse the channel and position attention information at the same time. Because the two can complement each other, make up for each other's shortcomings, and strengthen the model's representation capabilities.

3.3 Detection performance of different backbone networks and task networks with different attention modules

Given an input image, channel and position attention are used to focus on the "what" and "where" information, respectively, to enhance or suppress feature map information. To efficiently combine the two types of attention blocks, attention blocks can be assembled in a parallel or serial manner. a series of attention modules were conducted, such as CCP, CPC, and PCP, to explore the best combination of the two types of attention information. Table 4 lists the experimental results under the condition that the backbone networks are ResNet50 and ResNet101, and the task networks are Faster R-CNN, Cascade R-CNN, and Libra Faster R-CNN.

Effectiveness of Different Attention Modules: The parallel combination is better than the cascade combination method.

Taking R50 as the backbone network and Cascade R-CNN as the task network as an example (simplified to Cascade R-CNN-R50), compared with CCP and CPC, using PCP always achieves the best performance. Specifically, PCP achieves 2.81% and 2.93% improvements in mAP metrics, respectively. More importantly, for a specific facial posture category, the PCP-based Cascade R-CNN-R50 is able to achieve 89.67%, 87.57%, 87.57%, 86.40%, 84.56%, and 86.11% AP on the D-O, D-L, L-O, L-L, U-O, U-L, respectively, which is higher 2.52%, 1.91%, 3.41%, 2.01%, 4.81%, 2.2% AP than CCP-based Cascade R-CNN respectively. it was argued that the main reason is that it is easier to supplement the channel and position attention information in the parallel method, and the knowledge of the two dimensions extracted by the two branches will not interfere with each other during the feature extraction process but can complement each other during feature fusion. However, for the cascade manner, whether it is CCP or CPC, the information extracted by the attention of the previous stage is bound to affect the next stage. Taking the CCP module as an example, if there is a deviation in the channel attention information, further extracting the position attention based on the deviation information will definitely increase the deviation gap. For the combination of attendance information, for the Faster R-CNN model, it is more efficient to perform position attention first and then channel attention. On the contrary, when selecting the Cascade R-CNN and Libra Faster R-CNN as the task networks, first perform the channel attention outperforms the position attention. The reason may be that channel attention pays more attention to the category information of the detection box, and position attention can be used for improving the positioning accuracy of the box. For the calculation of the AP metric, the category accuracy takes precedence over the position accuracy, and it only makes sense to discuss the position accuracy when the category of the detection box is correct. So overall, the first use the channel attention outperforms the first use the position attention.

Table 4 Different task networks' performance under different attention modules

Model	Backbone	Attention	D-O	D-L	L-O	L-L	U-O	U-L	mAP
Faster R-CNN	R50	CCP	84.06	84.80	80.07	81.91	78.09	81.23	81.69
		CPC	**87.34**	82.99	80.60	82.00	79.88	81.10	82.32
		PCP	86.6	**86.98**	**86.26**	**86.67**	**83.03**	**86.28**	**85.97**
	R101	CCP	90.32	89.19	87.38	87.31	83.81	85.00	87.17
		CPC	91.20	89.34	87.82	86.98	83.55	85.36	87.38
		PCP	**91.55**	**90.36**	**90.10**	**90.05**	**85.96**	**87.92**	**89.32**
Cascade R-CNN	R50	CCP	87.15	85.66	84.16	84.39	79.75	83.91	84.17
		CPC	87.47	84.57	84.28	83.65	81.05	83.31	84.05
		PCP	**89.67**	**87.57**	**87.57**	**86.40**	**84.56**	**86.11**	**86.98**
	R101	CCP	**90.30**	87.48	**86.21**	**86.82**	81.40	83.41	85.94
		CPC	89.61	87.11	85.36	86.74	81.92	84.30	85.84
		PCP	90.21	**87.58**	86.10	86.57	**83.84**	**84.83**	**86.52**
Libra Faster R-CNN	R50	CCP	**88.76**	86.77	85.44	82.98	78.49	80.56	83.83
		CPC	**88.76**	86.05	**85.60**	83.13	**79.71**	79.07	83.72
		PCP	88.18	**87.25**	83.62	**85.58**	79.07	**83.94**	**84.61**
	R101	CCP	**89.19**	**88.77**	85.53	85.88	80.95	85.15	85.91
		CPC	88.11	87.15	86.21	85.74	81.10	**86.15**	85.80
		PCP	89.05	88.10	**87.55**	**86.71**	**81.80**	84.93	**86.36**

Note: CCP: Cascade Channel Position attention module; CPC: Cascade Position Channel attention module.

4 Visualization

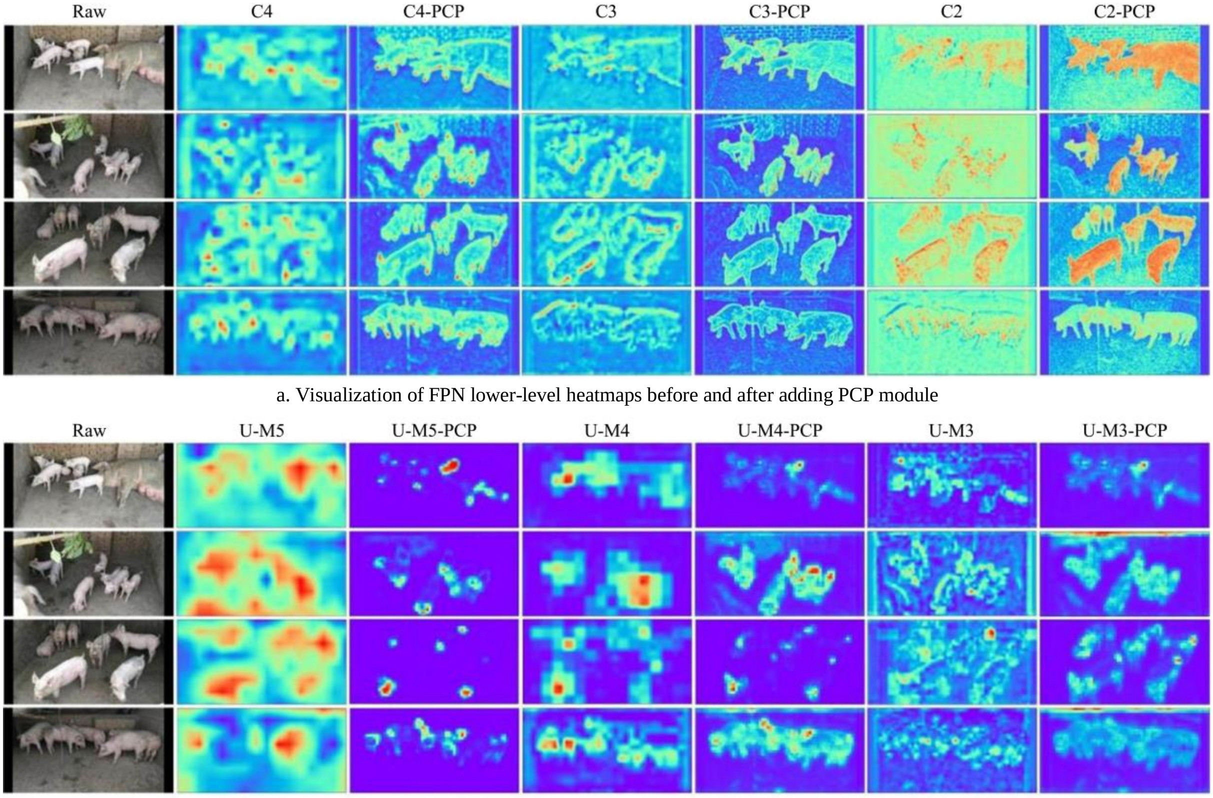
4.1 Visualization of heatmaps before and after adding PCP module

In order to understand the proposed PCP module more intuitively, with Faster R-CNN-R101 as an experimental model, the heatmaps of different stages before and after adding PCP module to FPN are visualized in Figure 9. For each input image, three channels were selected randomly for visualization. For the lower feature in the FPN structure (the bottom-up path in Figure 4), C2, C3, and C4 were chosen for visualization, after adding the PCP module, the corresponding results are C2-PCP, C3-PCP, C4-PCP, the corresponding visualization results are shown in Figure 9a. For higher feature (the top-down pathway in Figure 4), in order to ensure the same size as the corresponding lower dimensional feature maps, the output of M5, M4, and M3 was upsampled for visualization, denoting as U-M5, U-M4, and U-M3 The visualization results after adding the PCP module are marked as U-M5-PCP, U-M4-PCP, and U-M3-PCP, respectively, and the results are shown in Figure 9b.

For lower-feature heatmaps: In Figure 9a, for each input image, three of the channels (marked as #114 for C4, #138 for C3, and #252 for C2) were selected to display their corresponding heatmaps. It can be observed that although the task of this study was pig face pose detection, the lower-level heatmaps place more emphasis on the global body information of pigs. For different levels of the bottom-up path in FPN, from C2 to C4, the acquired heatmaps are gradually abstracted while paying more attention to detail. For example, the heatmap at stage C2 gives higher heat to the body part of the pig, while the heatmap at stage C4 puts more emphasis on the area of the hooves or ears. Compared with the non-PCP module, adopting the PCP module enables the aggregation of denser and richer contextual information in bottom-up paths at different levels. Especially, for C4 and C4-PCP, the PCP module strengthens the distinction between pig body and non-pig body parts while highlighting areas with rich biological information, such as hooves, thereby improving the accuracy of related tasks.

For higher-feature heatmaps: In order to visualize the heatmap of high-level features in FPN before and after adding the PCP module. Three channels were chosen in the top-down path of the FPN structure, marked as #58 for U-M5, #117 for U-M4, and #212 for U-M3. It could be observed that, compared with the heatmap of low-level features, the high-level features pay more attention to the areas related to the pig face, especially for the U-M5 and U-M4. The PCP module can capture clear semantic similarities and long-range distance relationships. For example, for the level of

U-M5, after the introduction of the PCP module, although the faces of different individuals are located in different positions of the image, the heatmaps can still be able to highlight most of the areas where the pig's face is. For the U-M5 and U-M4, the facial area information is particularly significant, and the number of faces can be clearly counted, which is also helpful for the task of counting group-housed pigs. Although some pigs are smaller (such as the second row of Figure 9b) or far away from the camera (such as the third row of Figure 9b), the PCP-based FPN is still able to detect the corresponding pig's face area. It should be noted that it was could not distinguish different face posture categories from the heatmaps obtained by FPN, the introduction of the PCP module can effectively filter the face area, and for the category of the face posture, it is left to the subsequent head network module (as show the Hi in Figure 3). In addition, after the introduction of the PCP module, there are great differences in the pig's face area concerned by different channels. Specifically, the U-M5-PCP pays more attention to the pig's nose position, while U-M4-PCP focuses on the pig's ear area. The position of the pig's nose is the basis for judging the upward, level, and downward postures, while the ear area is the basis for judging the posture of the lateral and head-on. In short, the facial semantic responses of pigs are significantly enhanced after adding the PCP module, and these visualizations further demonstrate the necessity of capturing long-range dependencies to improve feature representation for facial posture detection.



a. Visualization of FPN lower-level heatmaps before and after adding PCP module

b. Visualization of FPN higher-level heatmaps before and after adding PCP module

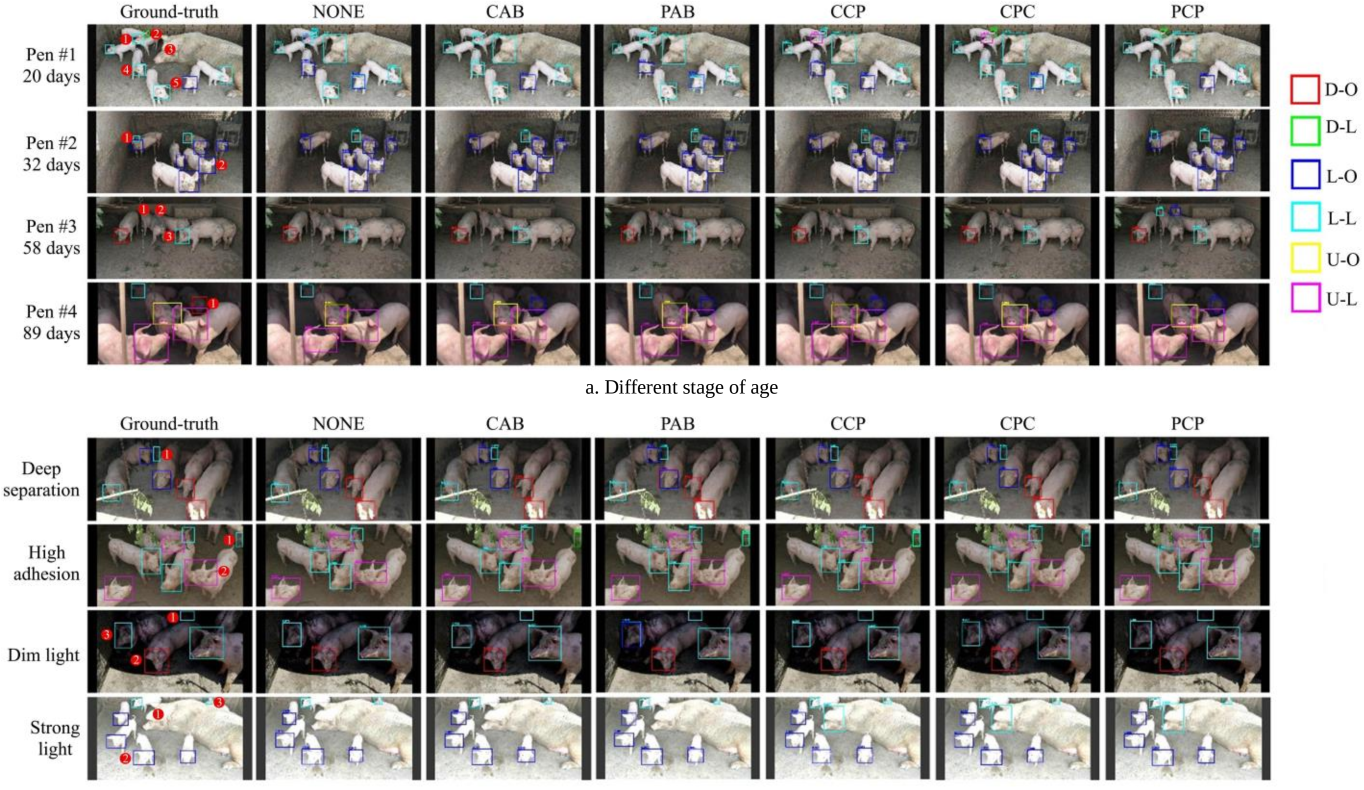
Note: Raw represents the input image; C4, C3, and C2 represent the different stages of the bottom-up pathway in Figure 4. C4-PCP, C3-PCP, and C2-PCP indicate the heatmaps after adding the PCP module to C4, C3, and C2. U-M5, U-M4, and U-M3 represent the different stages of the top-down pathway in Figure 4 after the applied upsample operation. U-M5-PCP, U-M4-PCP, and U-M3-PCP indicate the heatmaps after adding the PCP module to U-M5, U-M4, and U-M3.

Figure 9 Visualization of FPN heatmaps before and after adding PCP module

4.2 Visualization of predictions for different ages and scenarios

To further verify the effectiveness of different attention blocks or modules. Faster R-CNN-R101 was used as the experimental model. The test set is divided into different test subsets from different angles, and the main criteria for the division are the age of the pigs, the degree of adhesion between pigs, and the light intensity of the pigpen. For different stages of age, four pigpens in different age stages were selected as study subjects and visualized the facial posture detection results in Figure 10a. According to the degree of adhesion of the pigs, the pigs were divided into deep separation and high adhesion, and selected two samples to visualize the results. The corresponding results are shown in the first and second rows of Figure 10b. For the light intensity of the pigpen, it was divided into two scenarios in this study: dim light and strong light, and selected two samples to visualize the results as shown in the third and fourth rows of Figure 10b.

Qualitative Evaluation: 1) After adding attention information, it can improve the prediction confidence and position detection accuracy for pigs of different ages and life scenarios; 2) CPC and PCP can be used to improve the correct facial position detection but face posture category incorrect situation. For example, for the pigpen #1 number ②, the PAB and CCP modules predict its corresponding category as L-L, but the actual category is D-L, and the CPC and PCP modules can correctly predict it; 3) For facial areas that are far away from the camera or only partially visible, such as the pigpen #2 number ① and pigpen #3 number ① and ②, although the PCP module has lower confidence in the prediction of these face areas, at least facial posture can be detected correctly. 4) For the high adhesion scenario (the second row of Figure 10b numbered ①), due to the influence of the debris information, the pig face information cannot be detected for the model without attention blocks or modules. The CAB block can only detect the pig area, but the prediction of its category is wrong. CPC and PCP can eliminate redundant category box information and retain the category with higher confidence and position accuracy. At the same time, compared with CPC, the PCP module has higher confidence in the prediction of the face position, and the detection box is also more accurate. 5) For the strong light scenario (numbered ① in the last row of Figure 10b), the face area of the sow was not manually annotated, but the modules CCP, CPC, and PCP which simultaneously introduce channel and position attention information can further detect the corresponding areas, and the PCP can achieve the highest confidence. This indicates that the two kinds of attention information are fused at the same time, which helps to assist in labeling pig face posture detection data.



a. Different stage of age

b. Different degree of adhesion and light intensity

Note: Identify the different areas of attention prediction by input image number ①~⑤. Ground-truth represents the results of manual labeling. NONE represents the Faster R-CNN-R101 model that predicts results without any attention to information. CAB, PAB, CCP, CPC, and PCP represent the result of Faster R-CNN-R101 with channel attention block, position attention block, cascade channel position module, cascade position channel module, and parallel channel position module, respectively.

Figure 10 Visualization of predictions for different ages and scenarios

5 Conclusions

A feature pyramid network was proposed that simultaneously fuses channel and position attention information for pig face posture detection, adaptively capturing channel dependencies and the pixel-level long-range relationship in a parallel manner. First, we explore the performance of different models under the combination of three task networks and two backbone networks. Next, channel and position attention blocks are constructed to explicitly encode the interdependencies between channels and locations. Ablation experiments show that the position attention block is slightly inferior to the channel attention block. Furthermore, in this study, multiple ways were explored to effectively combine the two types of attention mechanisms, demonstrating that parallel permutations can provide better performance than the cascade manner. Subsequently, the heatmaps at different levels were visualized before and after adding the PCP module in the FPN to verify the proposed PCP module enhances the feature representation. Finally, the prediction results for different stages of age were visualized, and different degrees of adhesion, and light intensity, confirmed that the PCP module is the most robust.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 31671571), the Shanxi Province Basic Research Program Project (Free Exploration) (Grant No. 20210302124523; 20210302123408; 202103021224149; 202103021223141) and the Youth Agricultural Science and Technology Innovation Fund of Shanxi Agricultural University (Grant No. 2019027).

[References]

[1] Hansen M F, Smith M L, Smith L N, Salter M G, Baxte E M, Farish M, et al. Towards on-farm pig face recognition using convolutional neural networks. Computers in Industry, 2018; 98: 145–152.

[2] Marsot M, Mei J Q, Shan X C, Ye L Y, Feng P, Yan X J, et al. An adaptive pig face recognition approach using Convolutional Neural Networks. Computers and Electronics in Agriculture,2020; 173: 105386. doi: 10.1016/j.compag.2020.105386.

[3] Zhang K F, Li D, Huang J Y, Chen Y F. Automated video behavior recognition of pigs using two-stream convolutional networks. Sensors, 2020; 20(4): 1085. doi: 10.3390/s20041085.

[4] Condotta, I C, Brown-Brandl T M, Silva-Miranda K O, Stinn J P. Evaluation of a depth sensor for mass estimation of growing and finishing pigs. Biosystems Engineering, 2018; 173: 11–18.

[5] Valletta J J, Torney C, Kings M, Thornton A, Madde J. Applications of machine learning in animal behavior studies. Animal Behavior, 2017; 124: 203–220.

[6] Nasirahmadi A, Sturm B, Olsson A C, Jeppsson K H, Müller S, Edwards S, et al. Automatic scoring of lateral and sternal lying posture in grouped pigs using image processing and Support Vector Machine. Computers and electronics in agriculture, 2019; 156: 475–481.

[7] Shi C, Zhang J L, Teng G H. Mobile measuring system based on LabVIEW for pig body components estimation in a large-scale farm. Computers and electronics in agriculture, 2019; 156: 399–405.

[8] Chen C, Zhu W X, Liu D, Steibel J, Siegford J, Wurtz K, et al. Detection of aggressive behaviors in pigs using a RealSence depth sensor. Computers and Electronics in Agriculture, 2019; 166, 105003. doi: 10.1016/j.compag.2019.105003.

[9] da Fonseca F N, Abe J M, de Alencar Nääs I, da Silva Cordeiro A F, do Amaral F V, Ungaro H C. Automatic prediction of stress in piglets (Sus Scrofa) using infrared skin temperature. Computers and Electronics in Agriculture, 2020; 168: 105148. doi: 10.1016/j.compag.2019.105148.

[10] Sun L Q, Chen S H, Liu T, Liu C H, Liu Y. Pig target tracking algorithm based on multi-channel color feature fusion. Int J Agric & Biol Eng, 2020; 13(3): 180–185.

[11] Wu X W, Sahoo D, Hoi S C H. Recent advances in deep learning for object detection. Neurocomputing, 2020; 396: 39–64.

[12] Zhang Y Q, Chu J, Leng L, Miao J. Mask-refined R-CNN: A network for refining object details in instance segmentation. Sensors, 2020; 20(4), 1010. doi: 10.3390/s20041010.

[13] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 2018; 147: 70–90.

[14] Tassinari P, Bovo M, Benni S, Franzoni S, Poggi M, Mammi L M E, et al. A computer vision approach based on deep learning for the detection of dairy cows in free stall barn. Computers and Electronics in Agriculture, 2021; 182: 106030. doi: 10.1016/j.compag.2021.106030.

[15] Jiang M, Rao Y, Zhang J Y, Shen Y M. Automatic behavior recognition of group-housed goats using deep learning. Computers and Electronics in Agriculture, 2020; 177: 105706. doi: 10.1016/j.compag.2020.105706.

[16] Su D, Qiao Y, Kong H, Sukkarieh S. Real time detection of inter-row ryegrass in wheat farms using deep learning. Biosystems Engineering, 2021; 204: 198–211.

[17] Chen C, Zhu W, Steibel J, Siegford J, Wurtz K, Han J, et al. Recognition of aggressive episodes of pigs based on convolutional neural network and long short-term memory. Computers and Electronics in Agriculture, 2020; 169: 105166. doi: 10.1016/j.compag.2019.105166.

[18] Chen C, Zhu W X, Steibel J P, Siegford J M, Han J, Norton T J. Classification of drinking and drinker-playing in pigs by a video-based deep learning method. Biosystems Engineering, 2020; 196: 1–14.

[19] Yang A, Huang H, Yang X, Li S, Chen C, Gan H, et al. Automated video analysis of sow nursing behavior based on fully convolutional network and oriented optical flow. Computers and Electronics in Agriculture, 2019; 167: 105048. doi: 10.1016/j.compag.2019.105048.

[20] Hu Z W, Yang H, Lou T T, Hu G, Xie Q Q, Huang J M. Extraction of pig contour based on fully convolutional networks. Journal of South China Agricultural University 2018; 39(6): 111–119. (in Chinese)

[21] Hu Z W, Yang H, Lou T T. Dual attention-guided feature pyramid network for instance segmentation of group pigs. Computers and Electronics in Agriculture, 2021; 186: 106140. doi: 10.1016/j.compag.2021.106140.

[22] Tian M X, Guo H, Chen H, Wang Q, Long C J, Ma Y H. Automated pig counting using deep learning. Computers and Electronics in Agriculture, 2019; 163 104840. doi: 10.1016/j.compag.2019.05.049.

[23] Nasirahmadi A, Sturm B, Edwards S, Jeppsson K H, Olsson A C, Müller S, et al. Deep learning and machine vision approaches for posture detection of individual pigs. Sensors, 2019; 19(17): 3738. doi: 10.3390/s19173738.

[24] Zheng C, Zhu X, Yang X, Wang L, Tu S, Xue Y. Automatic recognition of lactating sow postures from depth images by deep learning detector. Computers and Electronics in Agriculture, 2018; 147: 51–63.

[25] Tong W, Chen W T, Han W, Li X J, Wang L Z. Channel-attention-based DenseNet network for remote sensing image scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020; 13: 4121–4132.

[26] Huang G, Liu Z, Van Der Maaten L, Weinberger K Q. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2017; pp.4700–4708.

[27] Chen C F, Gong D H, Wang H, Li Z F, Wong K Y K. Learning spatial attention for face super-resolution. IEEE Transactions on Image Processing, 2020; 30: 1219–1231.

[28] Woo S, Park J, Lee J Y, Kweon I S. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), IEEE, 2018; pp.3–19.

[29] Park J, Woo S, Lee J Y, Kweon I S. Bam: Bottleneck attention module. arXiv, 2018; arXiv preprint arXiv:1807.06514.

[30] Roy A G, Navab N, Wachinger C. Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018; pp.421–429.

[31] Fu J, Liu J, Tian H J, Li Y, Bao Y J, Fang Z W, et al. Dual attention network for scene segmentation. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.3146–3154. doi: 10.1109/CVPR.2019.00326.

[32] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2014; pp.580–587.

[33] Girshick R. Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, 2015; pp.1440–1448.

[34] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016; 39(6): 1137–1149.

[35] Lin T, Dollar P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019; pp.936–944.

[36] Rezatofighi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019; pp.658–666.

[37] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018; pp.6154–6162.

[38] Pang J M, Chen K, Shi J P, Feng H J, Ouyang W L, Lin D H. Libra R-CNN: Towards balanced learning for object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2019; pp.821–830.

[39] He K M, Zhang X Y, Ren S Q, Sun J. Deep Residual Learning for Image Recognition. computer vision and pattern recognition. In: 2016 IEEE Conference on Computer Vision and Patter Recognition (CVPR), IEEE, 2016; pp.770–778.

[40] Wu T Y, Tang S, Zhang R, Cao J, Zhang Y D. Cgnet: A light-weight context guided network for semantic segmentation. IEEE Transactions on Image Processing, 2020; 30: 1169–1179.

[41] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2018; 7132–7141.

[42] Open-mmlab/mmdetection. Available: https://github.com/open-mmlab/mmdetection. Accessed on [2022-01-15].

[43] Visual Object Classes Challenge 2012 (VOC 2012). Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012. Accessed on [2022-01-15].