Concurrent channel and spatial attention in Fully Convolutional Network for individual pig image segmentation

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

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

Abstract: The separation of individual pigs from the pigpen scenes is crucial for precision farming, and the technology based on convolutional neural networks can provide a low-cost, non-contact, non-invasive method of pig image segmentation. However, two factors limit the development of this field. On the one hand, the individual pigs are easy to stick together, and the occlusion of debris such as pigpens can easily make the model misjudgment. On the other hand, manual labeling of group-raised pig data is time-consuming and labor-intensive and is prone to labeling errors. Therefore, it is urgent for an individual pig image segmentation model that can perform well in individual scenarios and can be easily migrated to a group-raised environment. In order to solve the above problems, taking individual pigs as research objects, an individual pig image segmentation dataset containing 2066 images was constructed, and a series of algorithms based on fully convolutional networks were proposed to solve the pig image segmentation problem. In order to capture the long-range dependencies and weaken the background information such as pigpens while enhancing the information of individual parts of pigs, the channel and spatial attention blocks were introduced into the best-performing decoders UNet and LinkNet. Experiments show that using ResNext50 as the encoder and Unet as the decoder as the basic model, adding two attention blocks at the same time achieves 98.30% and 96.71% on the F1 and IOU metrics, respectively. Compared with the model adding channel attention block alone, the two metrics are improved by 0.13% and 0.22%, respectively. The experiment of introducing channel and spatial attention alone shows that spatial attention is more effective than channel attention. Taking VGG16-LinkNet as an example, compared with channel attention, spatial attention improves the F1 and IOU metrics by 0.16% and 0.30%, respectively. Furthermore, the heatmap of the feature of different layers of the decoder after adding different attention information proves that with the increase of layers, the boundary of pig image segmentation is clearer. In order to verify the effectiveness of the individual pig image segmentation model in group-raised scenes, the transfer performance of the model is verified in three scenarios of high separation, deep adhesion, and pigpen occlusion. The experiments show that the segmentation results of adding attention information, especially the simultaneous fusion of channel and spatial attention blocks, are more refined and complete. The attention-based individual pig image segmentation model can be effectively transferred to the field of group-raised pigs and can provide a reference for its pre-segmentation.

Keywords: pig, image segmentation, Fully Convolutional Network (FCN), attention mechanism, channel and spatial attention
DOI: 10.25165/j.ijabe.20231601.6528

Citation: Hu Z W, Yang H, Lou T T, Yang H W. Concurrent channel and spatial attention in Fully Convolutional Network for individual pig image segmentation. Int J Agric & Biol Eng, 2023; 16(1): 232–242.

1 Introduction

In an intensive breeding environment, the continuous increase in the density of pig breeding greatly increases the risk of infection and makes it more difficult to prevent and control swine fever. The automatic and effective identification of individual pigs in the group breeding environment, the construction of personalized profiles for pigs, and the establishment of a breeding traceability system are of great significance to the precise management of pig farms. One of the key steps and classic problems is how to separate individual pigs from the group-raised scene. Image analysis technology based on machine vision can provide a low-cost, non-contact, non-invasive monitoring method for group-raised pigs. Accurate and rapid detection of individual pigs in images can help find abnormal behaviors of pigs, take timely counter measures, and reduce the incidence of diseases. However, objective factors such as complex light changes in the pigpen, pig adhesion, and rigid occlusion have brought great difficulties to the study of individual pigs. Therefore, fast and accurate detection of pig targets in all-weather and multi-interference scenarios is a key problem that needs to be solved urgently.

The research of pig individual segmentation based on machine vision has made great progress in many aspects. Traditional pig image segmentation methods are mainly divided into static image segmentation and dynamic image segmentation. Common static segmentation methods include threshold segmentation, edge detection segmentation, watershed transformation, and morphological segmentation. Guo et al.[1] proposed a multi-object extraction method from top-view group-housed pig images based on adaptive partitioning and multilevel thresholding segmentation. Xu et al.[2] adopted GrabCut[3] and watershed segmentation of target object calibration to get the target pig areas. Dynamic image segmentation methods include the optical flow method,

Received date: 2021-02-16 **Accepted date:** 2022-09-28
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; **Hongwen Yan**, PhD, Associate Professor, Email: yhwhxh@126.com.
***Corresponding author: Hua Yang**, PhD, Professor, research interest: agricultural informatization. College of Information Science and Engineer, Shanxi Agricultural University, Taigu 030801, Shanxi, China. Tel: +86-13466883256, Email: yanghua@sxau.edu.cn.

frame difference method, and background difference method. Ma et al.[4] proposed a method to construct pig view feature vectors using the rectangular aspect ratio and low-frequency Fourier coefficients of sticky pigs, which can automatically segment pigs from surveillance videos in pigpens. However, the above methods mainly have the following two challenges: 1) These need to artificially select a large number of feature points for feature modeling, but in practice, it is inevitable that the feature point selection may be wrong or missing due to human limitations. 2) These only consider simple scenes, that is, it is effective when there is a large difference visible to the naked eye between the foreground and background of pigs, but they do not carry out in more complex scenarios such as pig adhesion and debris occlusion, which are more suitable for actual production practice.

Deep learning has been proven to be the most promising solution for the efficient processing of images in different environments[5]. As one of the most representative deep learning techniques, convolutional neural networks (CNNs) has powerful feature extraction ability for images, and it has been widely used in classification[6-8], object detection[9-11], image segmentation[12,13], and other vision tasks[14,15]. In the field of individual pig research, CNNs have been used in areas such as pig counting[16,17], pig face recognition[18-22], multi-target tracking[23,24], pig detection[25-28], recognition of the behaviors of pigs[29,30], and other tasks[31-33]. As a type of CNNs, fully convolutional networks (FCNs)[34] are widely used in the field of semantic segmentation and have achieved good performance[35,36]. However, in terms of individual pig image segmentation, there is relatively little research due to the lack of high-quality datasets and the existence of objective factors such as pig self-adhesion and pig house shading. Psota et al.[37] introduced a new dataset and method for instance-level detection of group-raised pigs, further using a single FCN to detect the location and orientation of each animal. Yang et al.[38] proposed a staged approach combining FCNs and Ostu thresholding to segment the sow image from the top-view perspective. Hu et al.[39] proposed a novel FCN model based on the combination of VGG16[40] and UNet[41], and conduct experiments with different batch sizes to explore the effect of batch size on the segmentation effect. Yang et al.[42] used the spatiotemporal information of the sow's location behavior to accurately segment the sow by FCN, and dynamically calculate the udder area and the length of the piglet for automatic identification of nursing behavior based on the geometric characteristics of the sow. Yang et al.[43] proposed a novel method using FCN-based semantic segmentation to further exploit the spatio-temporal relationship between the nursing sow and piglets. In addition to semantic segmentation, many scholars are also working on pig instance segmentation that can distinguish each individual pig. Hu et al.[44] proposed a dual attention-guided feature pyramid network and embed it into Mask R-CNN[45], Cascade Mask R-CNN[46], and HTC[47] to perform instance segmentation for group-raised pigs. Tu et al.[48] explored a new instance segmentation method based on the Mask R-CNN and soft-NMS[49] for adhesive group-housed pig images. However, the above method has the following two challenges: 1) Different parts of individual pigs have different contributions to the segmentation results. For example, pig trotters belong to pig parts, which are helpful for segmentation tasks, but pig manure or pigpen that do not belong to individual pigs should be eliminated from the model. In many cases, pig trotters are often mixed with pig manure, but the above segmentation models do not propose a special mechanism to distinguish between the two. 2) The growing environment of pigs is complex and changeable, and easy to stick and be blocked by sundries such as pigpens. However, the above methods mainly focus on group-raised pigs, and manual labeling of the pigs in the group-raising environment is time-consuming and labor-intensive. In addition, due to the adhesion of pigs, problems such as manual labeling errors are also prone to occur.

To address the first challenge, the attention mechanism increases the weight of regional information that is beneficial to the task, suppresses secondary information to improve the model effect, and has achieved good results in image segmentation[50-52] and object detection tasks in the open field. Inspired by this, a series of attention blocks was proposed to encode long-range dependencies between features within the feature map. Specifically, a self-attention mechanism was introduced to capture feature dependencies in channel and spatial dimensions in FCNs, respectively. Furthermore, different attention blocks were added to decoders UNet and LinkNet[53] to explore the effectiveness of each attention block. To address the second challenge, only trained the model on the individual pigs' dataset, and test the model on the individual and group-raised pig datasets. In this experimental setting, on the one hand, the complexity of manually constructing a dataset can be reduced, and labeling errors caused by objective factors such as adhesion can be avoided. On the other hand, the robustness and transferability of the model can be fully verified.

Overall the contributions can be summarized as follows:

1) Different encoder and decoder structures were constructed through comparative experiments to explore the most suitable encoder-decoder combination for individual pig image segmentation;

2) A novel model was proposed that combines channel with spatial attention information to capture long-range dependencies and introduce ablation experiments to verify the effectiveness of each block individually. Experiments show that adding two attention blocks at the same time can achieve the best segmentation results;

3) The visualization shows the heatmaps of the pig activation areas generated when the decoder structure is UNet after adding three different attention blocks, which intuitively shows the effectiveness of the attention information;

4) The ResNext50-UNet model, which was trained and performed best in the individual pig scenario, was transferred to the group-raised pig environment, and the segmentation results in different scenarios verified the robustness and transfer performance of the model.

2 Material and methods

2.1 Dataset Overview

2.1.1 Dataset Description

The experimental data of the study consisted of two aspects, one is individual pigs used for training models (denote as IND-Pig) and the other is group-raised pigs used for testing the models (denote as GRO-Pig). The two datasets were collected differently and played different roles in the overall model training. Considering that the adhesion of group-raised pigs brings great challenges to data labeling, only the IND-Pig dataset was labeled and used for model training. But for GRO-Pig dataset, was only used for model robustness testing. This method of training model

only on individual pigs and testing on group-raised pigs has the following two advantages:

1) Compared with labeling all the data, labeling only individual pigs can greatly reduce the labeling workload, while ensuring labeling accuracy, avoiding the problem of blurred labeling caused by the adhesion of pigs.

2) This method can be used to test the robustness of the model because the dataset cannot cover all possible scenarios, and only a more robust model is the key factor for the next application.

2.1.2 Individual pigs dataset

The IND-Pig dataset images were provided by JDD-2017 JD Finance Global Data Explorer Competition. This dataset included a total of 30 videos, where each video corresponded to only one Landrace pig in the pigpen scene, and each video was about 1 min long. Under the premise of ensuring that there is only one individual pig in an image, randomly intercept each frame from each video, and obtain a total of 1033 initial images with a resolution size of 1280×720 pixels. In order to obtain more diverse data, the following two steps were performed to preprocess operations to obtain the data-augmented dataset.

1) In order to adapt to the input of subsequent models, edge pixel padding is performed on frames cut out from the video, specifically, on the premise of ensuring the aspect ratio of 2:1, white pixels are filled around the image, and the image resolution changes from 1280×720 pixels to 1024×512 pixels. The whole process is shown in Figures 1a and 1b. In order to reduce the amount of model calculation and reduce the training video memory usage, an overall scaling operation was performed on the 1024×512 images, and finally, obtain 1033 images with a resolution of 512×256 pixels. The process is also shown in Figures 1a and 1b.

2) In order to enrich the dataset and improve the generalization ability of the models, data augmentation operations were performed on the first step of processed images. For each image, one to four augmentation operations are performed with a certain probability. Specifically, it is possible to perform the following three augmentation operations: flips 180° with a 50% probability value, adds Gaussian noise with a 50% probability value, changes the brightness with a 50% probability value, and randomly masks out a part of the image, in which the brightness value modification threshold is between 0.8-1.2, greater than 1 means dimming, and less than 1 means brightening. The width and height of the rectangular blocks of the random mask are any sizes between 50 and 100. The process is shown in Figures 1c and 1d.

After the above two steps, a total of 2066 images were acquired for the whole experiment, where 1346 were used for model training, and 308 and 412 were used for model validation and testing, respectively.

2.1.3 Group-raised pigs datasets

The group-raised experimental pigs' data were obtained from Jicun Town, Fenyang City, Shanxi Province, China, which was designated as JFS-Farm, and the Experimental Animal Management Center of Shanxi Agricultural University, China, designated as SXAU-Farm. The corresponding collection time, collection temperature, collection environment, and pigpen size of each pig farm are listed in Table 1. Canon 700D anti-shake camera was used to collect data, to meet the continuity of data, the data collection time for each pigpen is more than 60 min. Large white pigs, landrace pigs, and Duroc pigs were selected as the research objects. The age of pigs range from 20 d to 105 d, and each pigpen contained three to eight individual pigs. Finally, each pig farm selected four pigpens as the experimental objects, and a total of 45 pigs were obtained for model testing. Some different scene images in the GRO-Pig datasets are shown in Figure 2.

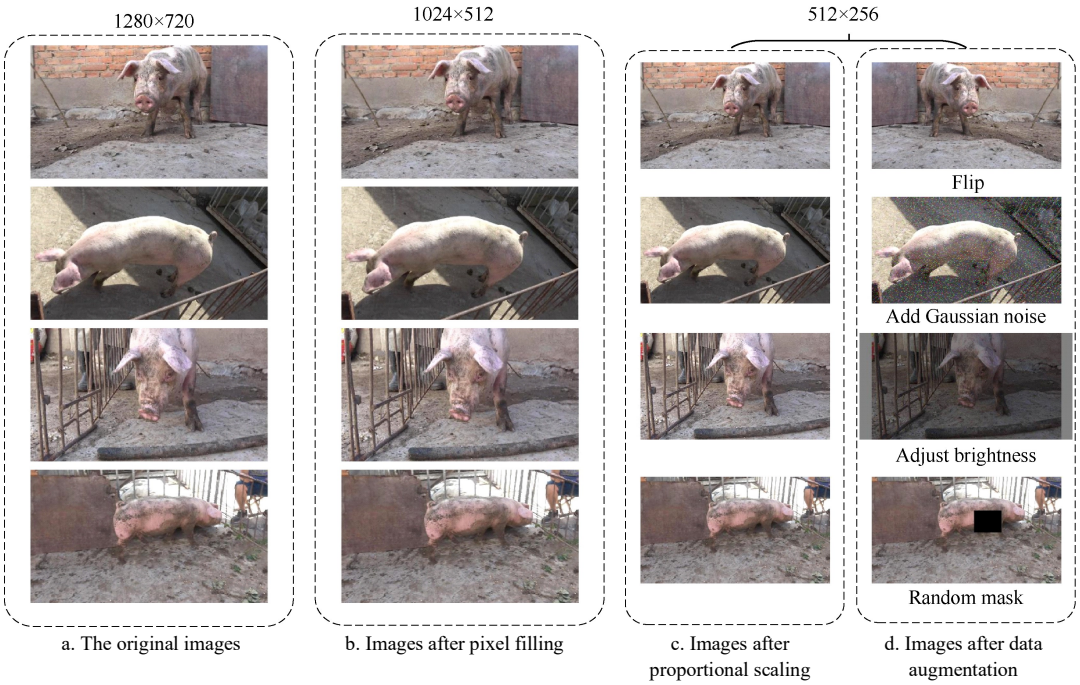


Figure 1 Preprocessing process of individual pig dataset

Table 1 Data collection environment information of different pig farms

Farm name	Collection time	Collection temperature	Collection environment	Pigpen size
JFS-Farm	June 1, 2019, 9:00–14:00	Sunny, 23°C–29°C	Outdoor, bright light	3.5 m×2.5 m×1 m
SXAU-Farm	October 13, 2019, 10:30–12:00	Cloudy, 10°C–19°C	Indoors, low light	4 m×2.7 m×1 m

Note: For the pigpen size, x×y×z means that the pigpen length is x, the width is y, and the height is z.

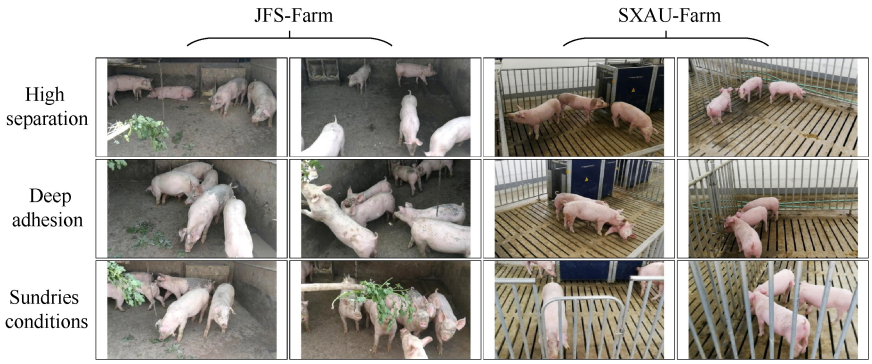


Figure 2 Group-raised dataset with different scenes

2.2 Individual pig image segmentation model

2.2.1 Model overview

Fully Convolutional Networks (FCN) architectures have been successfully applied in many fields, especially in semantic segmentation tasks, which have achieved state-of-the-art performance. The common semantic segmentation models based on FCN ideas are UNet (as shown in Figure 3a) and LinkNet (as shown in Figure 3c). Generally, FCN organizes the model through the encoder-decoder structure, the encoder can effectively capture the context information, while the decoder is more helpful in recovering the position content. The information exchange between the encoder and the decoder can be carried out by means of skip connections. By fusing the hierarchical features of the backbone, the encoder-decoder structure gradually increases the spatial resolution and restores the missing details. However, the traditional encoder-decoder structure has the following two weaknesses:

1) The local features of each layer in the encoder or decoder are independent and lack long-range dependencies information, which may lead to the misclassification of objects or stuff.

2) The encoder and decoder at the same layer simply perform linear stacking directly through skip connections, and without considering the nonlinear dependencies between feature maps.

To overcome the above two shortcomings, attention blocks were proposed to combine channel and spatial attention named channel attention block (CAB), spatial attention block (SAB), and concurrent channel and spatial attention block (CSA) to enhance feature information capture respectively. To verify the effectiveness of the corresponding attention blocks, CAB, SAB, and CSA blocks were embedded into the UNet (as shown in Figure 3b) and LinkNet (as shown in Figure 3d).

2.2.2 Channel Attention Block (CAB)

As each channel in the feature map can be regarded as a response to a specific category, and different semantic responses are related to each other. In order to make full use of the dependencies between channels to enhance the feature maps between interdependent channels and further improve the feature representation of specific semantics, a channel attention block (CAB) was built to explicitly model interdependencies between channels. The structure of the CAB is illustrated in Figure 4. The CAB block performs the following three steps on the input feature map **I** to obtain the output feature map $\mathbf{I}_{CAB}$ recalibrated by channel attention.

1) Different from the spatial attention block, to generate the channel attention aware features, two branches were conducted for the input feature $I \in R^{H\times W\times C}$, corresponding to generate two attention maps named **K** and **L**, which embedded the global spatial information via global average and max pooling operations, respectively, where $\{\mathbf{K}, \mathbf{L}\} \in R^{1\times 1\times C}$, H, and W represent the height and width of the feature map, C represents the channel numbers of the input feature I;

2) The two feature maps **K** and **L** are respectively dimensionally transformed by two fully connected layers, taking the K feature map as an example, the weights of the two fully connected layers are W_1 and W_2, where and $W_2 \in R^{\frac{C}{2}\times C}$. After the two fully connected layers are completed, the ReLU or sigmoid activation functions are used to perform nonlinear transformation operations to obtain the channel attention maps **M** and **N**, respectively, where $\{\mathbf{M}, \mathbf{N}\} \in R^{1\times 1\times C}$. Then pass the **M** and **N** together to produce R with mix attention map, which contains the interdependencies between channel maps;

3) Finally, perform an element-wise multiplication operation on the channel attention map R and the input feature map **I**, and the operation result with the original feature map **I** bitwise to get the final channel attention recalibrated feature map $\mathbf{I}_{CAB} \in R^{H\times W\times C}$.

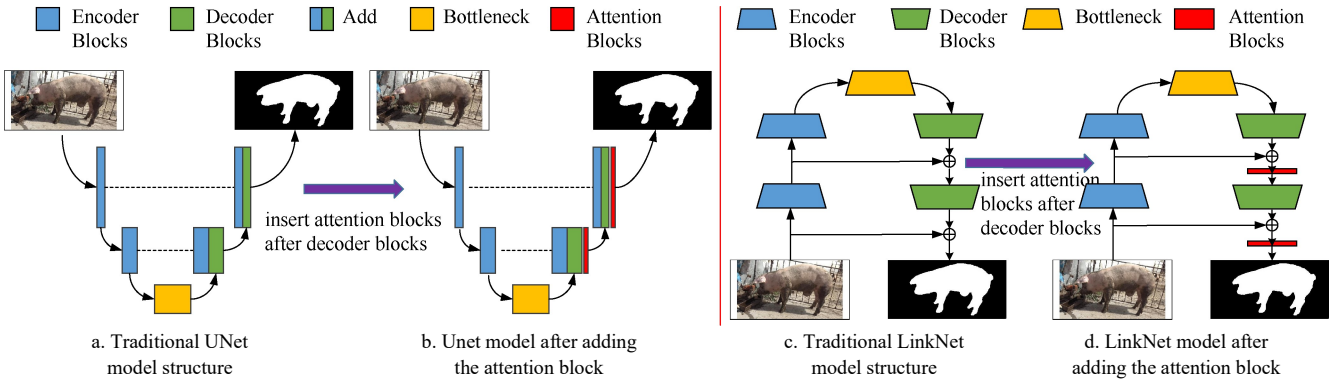
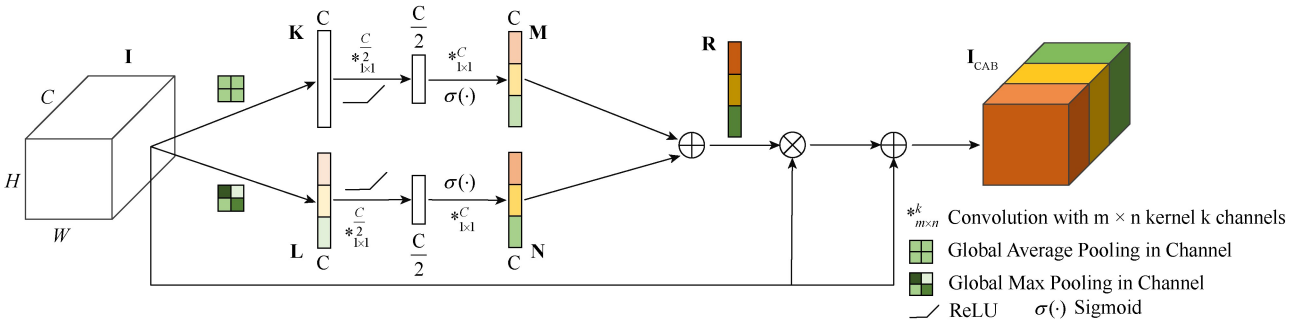


Figure 3 UNet and LinkNet models before and after adding the attention block



Note: *H*, *W*, *C* represent the height, width and channel numbers of the feature map **I**, {**K**, **L**} mean the feature maps, {**M**, **N**, **R**} denote the channel attention maps, $\mathbf{I}_{CAB}$ means the final channel attention recalibrated feature map, $\sigma(\cdot)$ means the sigmoid function.

Figure 4 The structure of channel attention block

The formulaic description of the above process is shown in Equations (1) and (2):

$$O_c(\mathrm{I})=\sigma\left(Full_{C\times\frac{C}{2}}\left(\delta\left(Full_{\frac{C}{2}\times C}\left(\mathbf{I}^c_{avg}\right)\right)\right)\right)+\sigma\left(Full_{C\times\frac{C}{2}}\left(\delta\left(Full_{\frac{C}{2}\times C}\left(\mathbf{I}^c_{max}\right)\right)\right)\right) \quad (1)$$

$$\mathbf{I}_{CAB}=\mathbf{I}+\mathbf{I}\times O_c \quad (2)$$

where, O_c represents the final merged channel attention map *R*, $\delta(\cdot)$ denotes the ReLU activation function, $\sigma(\cdot)$ denotes the sigmoid function, $\mathrm{Full}_{m\times n}(\cdot)$ represents the fully connected operation with m and *n* neuron nodes in the hidden and output layer. $\mathbf{I}^c_{avg}$ and $\mathbf{I}^c_{max}$ denote the global average *K* and max pooling *L* in channel dimension for input feature map **I**, and $\mathbf{I}_{CAB}$ represents the output of channel attention block.

2.2.3 Spatial Attention Block (SAB)

Discriminant feature representations are essential for individual pig image segmentation, different regions in the same feature map should be treated differently. For example, compared with pig noses, pigpens should be given lower attention. In order to generate dense, pixel-by-pixel context information to model rich contextual relationships over local features, a spatial attention block (SAB) was introduced as shown in Figure 5, which uses all pixels in a single feature map to weigh the response value of the target pixel to obtain a spatial attention map. The original feature map is guided by the attention map to select the location information to generate feature maps containing dense contextual associations. To generate spatial attention maps by exploiting the inter-spatial relationship in local feature maps. The spatial information can be excited while squeezing the channel features to encode a wide range of contextual information into local features. The entire SAB operation process includes the following three steps:

1) Given a local input feature map $\mathbf{I}\in R^{H\times W\times C}$, new feature maps were obtained **A**, **B**, and **D** through three different convolution operations, respectively, where $\{\mathbf{A}, \mathbf{B}, \mathbf{D}\}\in R^{H\times W\times 1}$. The feature map **A** is obtained by performing a convolution operation on **I** with a convolution kernel size of 1×1 and a channel number of 1. The sigmoid activation function is then used to rescale the eigenvalues between 0 and 1. Global average pooling and global max pooling were performed on the feature map **I** respectively in the channel dimension to obtain feature maps **B** and **D**, and then perform the element-wise summation operation to obtain the superimposed pooled feature map $\mathbf{M}\in R^{H\times W\times 1}$. In order to obtain the nonlinear representation of the feature map, a convolution operation with a convolution kernel size of 1×1 was performed on the feature map **M**, and also apply the sigmoid function for nonlinear representation was to generate the attention map $\mathbf{E}\in R^{H\times W\times 1}$.

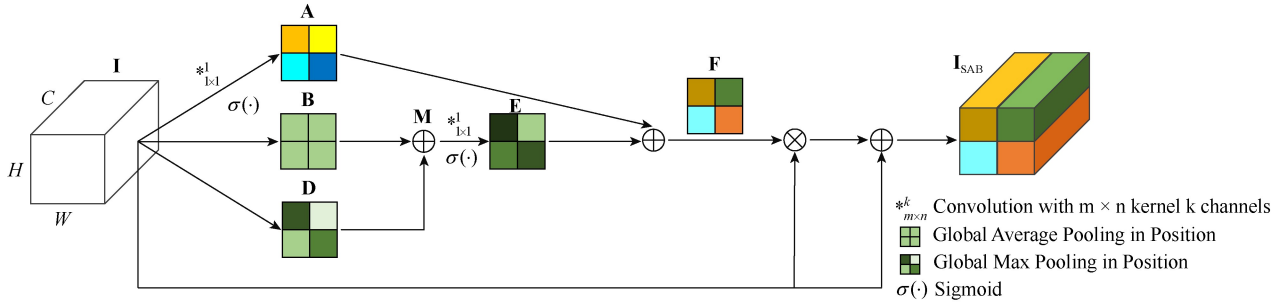
2) Attention maps **A** and **E** were combined to generate the final spatial attention map $\mathbf{F}\in R^{H\times W\times 1}$ to perform spatial information selection. Each element $(\cdot)_{i,j}$ of the feature maps $\{A_{i,j}, B_{i,j}, D_{i,j}, E_{i,j}, F_{i,j}\}$ represents the linearly combined representation for all channels for a spatial position (i, j).

3) After obtaining the attention map **F**, the element-wise multiplication was performed between **I** and **F**, and then residually concatenate the result of the addition with the original input feature map **I** to produce the spatial attention calibrated feature map $\mathbf{I}_{SAB}\in R^{H\times W\times C}$.

The formulaic description of the above process is shown in Equations (3) and (4):

$$O_s(\mathbf{I})=\sigma\left(Conv_{1\times1\times1}(\mathbf{I})\right)+\sigma\left(Conv_{1\times1\times1}\left(\mathbf{I}^s_{avg}+\mathbf{I}^s_{max}\right)\right) \quad (3)$$

$$\mathbf{I}_{SAB}=\mathbf{I}+\mathbf{I}\times O_s \quad (4)$$



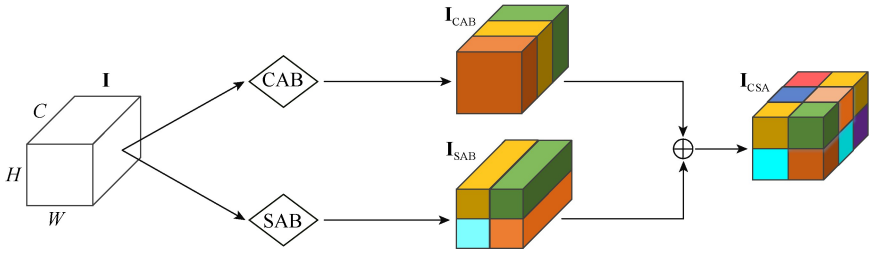
Note: *H*, *W*, and *C* represent the height, width, and channel numbers of the feature map **I**, {**A**, **B**, **D**} denote the feature maps through three different convolution operations, **M** means the superimposed pooled feature map, {**E**, **F**} represent the attention map, $\mathbf{I}_{SAB}$ means the spatial attention calibrated feature map.

Figure 5 The structure of spatial attention block

where, $O_s \in R^{H\times W\times 1}$ denotes the final spatial attention map **E**, $\mathrm{Conv}_{1\times1\times1}(\cdot)$ denotes the convolution operation with convolution kernel size 1×1 and the channel number is 1. The $\mathbf{I}^s_{avg} \in R^{H\times W\times 1}$ and $\mathbf{I}^s_{max} \in R^{H\times W\times 1}$ denote the global average and max polling operations in **I** respectively, also represent the *B* and *D* in Figure 6. The $\mathbf{I}_{SAB}$ represents the output of the spatial attention block.

2.2.4 Concurrent Channel and Spatial Attention (CSA)

In order to learn richer semantic information among different feature maps, two attention blocks were simultaneously introduced, CAB and SAB, to recalibrate the input feature map **I** in the channel and spatial dimensions, respectively, which helps the two kinds of attentional information to complement each other. The CAB can give differentiated information to different channels of the feature map, and increase the weight value of the channel including the pig area. SAB can give different weights to different position features in the same feature map. The combination of the two attention blocks on the one hand will make the channel and spatial information complements each other to further improve the accuracy of pig image segmentation boundaries. On the other hand, the recalibration of feature maps with two kinds of attention information can adequately capture long-range contextual information. In addition, more importantly, CAB and SAB blocks can be easily embedded into FCN-based models, with plug-and-play properties.



Note: *H*, *W*, and *C* represent the height, width, and channel numbers of the feature map **I**, $\mathbf{I}_{CAB}$and $\mathbf{I}_{SAB}$ mean the results of Channel Attention Block and Spatial Attention Block, $\mathbf{I}_{CSA}$ represents the recalibration of feature maps with two kinds of attention information.

Figure 6 The structure of concurrent channel and spatial attention block

3 Results and discussion

3.1 Experimental parameters and evaluation metrics

3.1.1 Experimental parameters

The experimental platform is configured with a 16 GB Tesla V100 GPU, the operating system is Ubuntu 16.04, and the code is written using the PyTorch framework. Adam[54] is used as the optimizer, the optimizer base learning rate hyperparameter is set to 0.001, and if there is no performance improvement within 10 epochs on the validation set, the learning rate size is modified to the current learning rate multiplied by 0.5. Besides, the mini-batch size was set to 32, that is, each batch trains 32 images, traversing all the training dataset is called one round of iteration, and the number of iteration rounds was set to 150, while most models will converge in about 100 epochs. In order to accelerate the model convergence speed, the weights pre-trained on ImageNet[55] were transferred as the initial weight information of the encoder models, such as MobileNetV2[57], VGG16, ResNet50[58], and ResNext50[59]. Inspired by Milletari[56], dice loss was adopted as the loss function for model training.

3.1.2 Evaluation Metrics

The harmonic mean of precision and recall (F1) and Intersection over-Union (IOU) were adopted, which are commonly used in the field of image segmentation, as evaluation metrics to measure the performance of the model for pig image segmentation, which can be formulated as shown in Equations (5) to (6).

$$P=\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fp}},\quad R=\frac{\mathrm{tp}}{\mathrm{tp}+\mathrm{fn}},\quad \mathrm{F1}=2\cdot\frac{P\cdot R}{P+R} \tag{5}$$

$$\mathrm{IOU}(p,q)=\frac{p\cap q}{p\cup q} \tag{6}$$

where, *P* and *R* represent Precision and Recall, respectively, Precision is the proportion of all positive predictions that are correct, Recall is the proportion of all real positive observations that are correct. The true positive (tp) denotes the number of pixels where both the true and the actual prediction are pigs. The false positive (fp) represents the number of pixels that are background but are actually predicted to be pigs. The false negative (fn) denotes the number of pixels that are pigs but are actually predicted to be background. IOU(*p*,*q*) means the intersection over union between the prediction results *p* and ground truth *q*, $p\cap q$ means the intersection area of *p* and *q*, $p\cup q$ means the union area of *p* and *q*.

3.2 Main results

3.2.1 Experimental results with different encoder and decoder structures

The encoder-decoder structure is usually used in image semantic segmentation tasks, the main function of the encoder is to extract image features, and the decoder is mainly to restore image features. Usually, the encoder and decoder can be replaced by various structures. MobileNetV2[57], VGG16, ResNet50[58], and ResNext50[59] were selected, which are widely used, as our encoder. For the decoder, UNet, LinkNet, FPN[60], and PSPNet[61] were chosen to recover feature maps of different resolutions extracted by the encoder. the performance of each combined model was evaluated on the test set by combining the encoder and decoder in pairs, and the results are listed in Table 2.

Same encoder with different decoders: Under the condition of the same encoder, using different decoders has a certain impact on the performance of the models. Compared with decoders FPN and PSPNet, the UNet and LinkNet perform better. Take model ResNet50-UNet (means choose ResNet50 as encoder and UNet as decoder, same as below) as an example, compared with using FPN and PSPNet as decoders, the IOU metric is improved by 0.7% and 2.06%. The same phenomenon happened to the F1 indicator, when the model is ResNext50-UNet, it achieves 98.22% in F1 metric, which outperforms other same encoder methods especially FPN and PSPNet by a large margin. In particular, compared with the FPN and PSPNet decoders, using UNet can improve 0.27% and 1.53% in F1 metric when selecting ResNext50 as the encoder. In addition, the performance of the UNet and LinkNet decoders are closer, the reason why this happens may be that both of these two decoders introduce the skip connection during the decoding process, which enables high-level features to perceive the existence of low-level features. Because high-level features pay more

attention to semantic information, and low-level features are more conducive to the acquisition of position information, the combination of the high and low features is conducive to ensuring the accuracy of position content while restoring semantic information.

Same decoder with different encoders: For the case where the decoder is the same but with different encoders, the experimental results are significantly diverse. Especially, the ResNext50-based encoder generally achieves the best F1 and IOU scores, although ResNext50 is not the best performer when the decoder is selected as PSPNet, it still produces competitive results. Take UNet as the decoder as an example, ResNext50-based encoder yields 98.22% and 96.57% in F1 and IOU metrics, compared with the MobileNetV2-based encoder, which brings 0.54% and 0.98% improvement respectively. Meanwhile, compared with the MobileNetV2-LinkNet, ResNext50-LinkNet can improve the segmentation performance by 0.72% and 1.3% in F1 and IOU metrics. Furthermore, the ResNet50 and ResNext50-based encoders can reach better performances than the encoders MobileNetV2 and VGG16, the main reason is that the two network structures of ResNet50 and ResNext50 are deeper, and the extracted feature information is more abundant, which can refine the edge area of pig image segmentation, further help to improve the segmentation accuracy. From the experimental results, a deeper network may be more conducive to the realization of the pig image segmentation task.

3.2.2 Ablation study for different Attention Blocks

In order to explore the effect of channel and spatial attention blocks on the performance of pig image segmentation, MobileNetV2, VGG16, ResNet50, and ResNext50 were chosen as the encoder, and the two best performance models UNet and LinkNet as the decoder, respectively. Under the same experimental conditions, the channel attention block CAB, the spatial attention block SAB and CSA that simultaneously fuse channel and spatial attention are compared. We also conduct comparative experiments with existing attention modules CBAM[52], BAM[62], and SCSE[51]. The experimental results are listed in Table 3.

Table 2 Evaluation performance in test set with different encoder and decoder structures

Encoder	Decoder	F1/%	IOU/%	Encoder	Decoder	F1/%	IOU/%
MobileNetV2	UNet	**97.68**	**95.59**	VGG16	UNet	**97.71**	**95.70**
	LinkNet	97.28	94.86		LinkNet	97.36	95.10
	FPN	97.50	95.27		FPN	96.94	94.40
	PSPNet	94.49	89.83		PSPNet	96.10	92.94
ResNet50	UNet	**97.95**	**96.15**	ResNext50	UNet	**98.22**	**96.57**
	LinkNet	97.74	95.78		LinkNet	98.00	96.16
	FPN	97.59	95.45		FPN	97.95	96.06
	PSPNet	96.86	94.09		PSPNet	96.69	93.87

Note: The bold font means the best results under the same encoder but with different decoder conditions. F1: The harmonic mean of Precision and Recall; IOU: Intersection over-Union.

Table 3 Ablation study of different attention blocks with UNet and LinkNet encoder

Encoder	Decoder	Block	F1/%	IOU/%	Decoder	Block	F1/%	IOU/%
MobileNetV2	UNet	None	97.68	95.59	LinkNet	None	97.28	94.86
		CBAM	97.78	95.72		CBAM	97.41	95.01
		BAM	97.83	95.89		BAM	97.52	95.12
		SCSE	97.91	96.02		SCSE	97.59	95.38
		CAB	97.89	95.98		CAB	97.32	94.95
		SAB	97.88	95.97		SAB	**97.70**	**95.57**
		CSA	**97.99**	**96.13**		CSA	97.65	95.55
VGG16	UNet	None	97.71	95.70	LinkNet	None	97.36	95.10
		CBAM	97.62	95.74		CBAM	97.43	95.18
		BAM	97.70	95.81		BAM	97.46	95.24
		SCSE	97.75	95.88		SCSE	97.52	95.33
		CAB	97.61	95.59		CAB	97.29	94.99
		SAB	97.72	95.73		SAB	97.45	95.29
		CSA	**97.83**	**95.95**		CSA	**97.64**	**95.59**
ResNet50	UNet	None	97.95	96.15	LinkNet	None	97.74	95.78
		CBAM	97.98	96.19		CBAM	97.79	95.93
		BAM	98.00	96.23		BAM	97.83	95.98
		SCSE	98.03	96.25		SCSE	97.89	96.03
		CAB	98.02	96.18		CAB	97.95	96.10
		SAB	**98.06**	**96.30**		SAB	97.82	95.91
		CSA	98.04	96.29		CSA	**97.99**	**96.17**
ResNext50	UNet	None	98.22	96.57	LinkNet	None	98.00	96.16
		CBAM	98.24	96.59		CBAM	98.03	96.17
		BAM	98.25	96.62		BAM	98.07	96.27
		SCSE	98.25	96.66		SCSE	98.14	96.33
		CAB	98.17	96.49		CAB	97.93	96.13
		SAB	98.27	96.64		SAB	98.04	96.25
		CSA	**98.30**	**96.71**		CSA	**98.21**	**96.55**

Comparison with no attention block: After adding the attention blocks, the segmentation metrics have been improved to varying degrees, especially the SAB and CSA attention blocks improve performance remarkably under the same experimental conditions. Take the encoder as VGG16 and the decoder as LinkNet as an example, after adding SAB and CSA blocks to the decoder, compared to the baseline without attention blocks, SAB and CSA increase by 0.09% and 0.28% respectively in F1 metric, and 0.19% and 0.49% in the IOU metric respectively. In addition, when selecting MobileNetV2 as the encoder and UNet as the decoder, after adding the CSA block, the value of F1 changes from 97.68% to 97.99%, about a 0.31% absolute improvement. When the SAB block is embedded in the decoder, on the IOU metric, the absolute increase is 0.54%. The above results prove the effectiveness of the attention block in the pig image segmentation task, which is mainly due to the fact that the attention mechanism can apply

higher attention information to the parts that are beneficial to the pig area so that the models pay more attention to the acquisition of regional information that is beneficial to the task.

Comparison of different attention blocks: Compared with CAB block, the SAB block can generally achieve better performance, although the improvement is limited. For the ResNext50-UNet model, which chooses the ResNext50 as the encoder and UNet as the decoder, compared to the introduced CAB, applying SAB to increase 0.10% and 0.15% respectively in F1 and IOU metrics. The reason lies that channel attention can give differentiated information to different channels of the feature map, and increase the weight value of the channel including the parts of the pig region. Compared with channel attention, spatial attention is more fine-grained and can distinguish the pig area and the background more finely. The channel attention will treat the feature maps in the same channel as having equal importance, which affects the performance to a certain extent. In addition, after introducing CSA for channel and spatial attention at the same time, the problem of missing attention extraction caused by using one of the attention information alone can be further improved. Although on some models (such as MobileNetV2-LinkNet and ResNet50-UNet), adding the CSA block does not achieve the best results, the performance is still competitive compared with the best results. Take ResNext50-UNet as an example, attended CSA block yields to 98.30% and 96.71% respectively in F1 and IOU, and achieved the best results. The performance difference can be explained that the proposed CSA block establishing the complementary relationship between SAB and CAB at different scales feature maps, resulting in more elaborated and semantically abstract representations, which are generally applicable to extract the global perception information to further achieve better segmentation results.

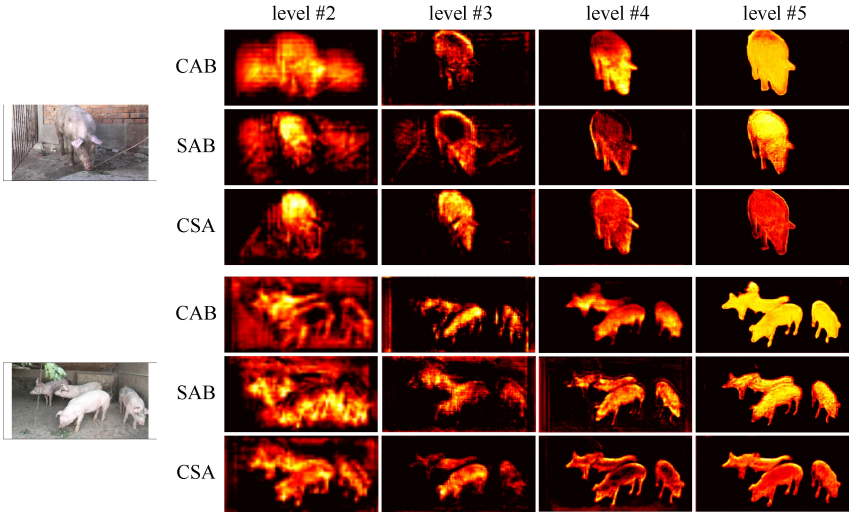
Comparison of existing attention blocks: Compared with existing CBAM, BAM, and SCSE modules that incorporate both channel and spatial attention information, our CSA attention module achieves better results under various encoder and decoder combinations. Specifically, select VGG16 as the encoder, LinkNet as the decoder, adding the CSA attention block improves the F1 and IOU values by 0.21%, 0.18%, 0.12%, and 0.41%, 0.35%, 0.26%, respectively compared to CBAM, BAM, and SCSE attention blocks. This proves that our attention block is well suited for the individual pig image segmentation task. Furthermore, for three existing attentions, CBAM, BAM, and SCSE, SCSE outperforms BAM, and BAM outperforms CBAM.

3.3 Visualization results

3.3.1 Visualization of different attention block feature maps at different layers

In order to more intuitively understand the effectiveness of the attention mechanism, taking the best performance ResNext50-UNet model as an example, the CAB, SAB, and CSA attention blocks were added to the model respectively and visualized the feature maps of different layers filtered by the attention information. In order to verify the robustness of the model, an image from the JFS-Farm and SXAU-Farm datasets was selected for visual display. The corresponding results are shown in Figure 7.

Compared with CAB and SAB attention blocks, the response to pig semantic areas is more noticeable after adding CSA block. The CSA ensemble network can well encode the information in the pig pixel region and aggregate key features, thereby obtaining smoother segmentation boundaries and higher heat values. Furthermore, for adding channels or spatial attention individually, compared with CAB blocks, the introduction of SAB block can aggregate denser and richer contextual information. After adding SAB, the segmentation boundary is clearer, which can better distinguish the pig area from the background. In addition, no matter which attention block is used, with the deepening of the decoder layers, the segmentation outline of individual pigs becomes clearer. Specifically, at a shallower level, the outline of the pig is only roughly segmented, and even difficult to distinguish whether it is a complete pig individual, but with the deepening of the decoder layer, even for pig trotters, pig mouths, pig eras, and other parts, the attention block is still able to get better segmentation. It is worth noting that only trained the model on the individual pig dataset, but when visualized the feature maps of group-raised environment pigs with the same attention, it still showed similar patterns. This fully proves that the proposed CAB, SAB, and CSA in this study shave strong robustness and can be applied to more complex application scenarios.



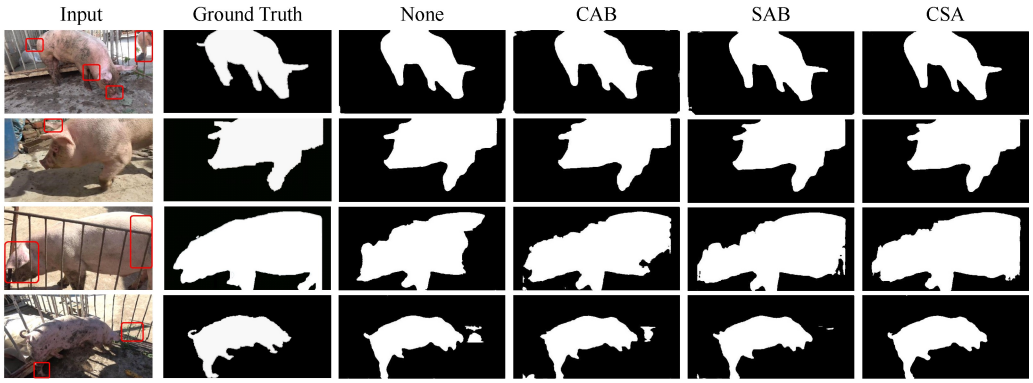
Note: The UNet decoder contains five layers the bottom layer denotes level #1, and the top defines level #5. The red regions are assigned more weight while the black areas are given less attention.

Figure 7 Visualization of feature maps with different attentions at different layers after attention filtering

3.3.2 Visualization of prediction results

In order to further verify the robustness of different attention blocks in different scenarios, ResNext50-UNet was chosen as the base model, Figure 8 and Figure 9 provide the qualitative visualization results of adding in different scenarios for individual pigs and group-raised pigs, respectively. In addition, the

individual pig dataset mainly includes three scenarios: normal, occlusion by debris such as pigpen, and uneven lighting condition (note that in many cases, multiple scenarios may be mixed, for image with multiple scenes at the same time, it is artificially classified according to the tendency). The statistical results of IOU metric of the ResNext50-Unet model adding different attention blocks in each scenario are listed in Table 4. It should be noted that in order to test the adaptability of the model, only trained the model on the dataset of a single pig, and the data corresponding to group-raised pigs did not participate in the training process of the model. In addition, red boxes are used to mark areas that are prone to mis-segmentation and challenges.



Note: The first to fourth rows represent pigs with similar regions (human legs), incomplete individual information, occlusion by debris, and uneven lighting conditions.

Figure 8 Visualization results of ResNext50-UNet with different attention blocks on different scenarios in the individual pig dataset

Note: Rows 1 and 2, Rows 3 and 4, and Rows 5 and 6 represent group pigs under high separation, deep adhesion, and pigpen occlusion conditions, respectively.

Figure 9 Visualization results of ResNext50-UNet with different attention blocks on different scenarios in the group-raised pig dataset

Table 4 IOU results of ResNext50-UNet with different attention blocks on different scenarios in the individual pig dataset

Scene	Block	IOU%	Scene	Block	IOU%	Scene	Block	IOU%
One	None	97.21	Two	None	95.28	Three	None	95.96
	CAB	97.17		CAB	95.19		CAB	96.02
	SAB	97.62		SAB	95.42		SAB	96.14
	CSA	**97.73**		CSA	**95.57**		CSA	**96.21**

Note: One represents the normal individual scene, Two represents the occlusion by debris scene, and Three represents the uneven lighting conditions scene.

For individual dataset: The use of attention blocks especially the CSA block, can correctly segment difficult parts, for example, for scenes occluded by debris. For example, for the scene where the sundries are occluded (the third and fourth rows in Figure 8), even if pigpen divides the individual pigs, the introduction of CSA can still be more accurately segmented. Although in some cases the model with attention block only slightly improves the segmentation performance (the first and second rows in Figure 8), in terms of detail processing, the edge prediction of the attention-based model is smoother and the pig outline is more complete. In addition, according to Table 4, for more complex scenes, the effect of light intensity on the results is weaker than the scene with debris such as pigpens. The reason lies in that in the preprocessing part, the data augmentation operation on the intensity of light is introduced, so that the model can learn the knowledge of light intensity. Channel and spatial attention can selectively capture contextual information, which in turn significantly improves semantic segmentation consistency.

For group-raised dataset: Although only an individual pig dataset was used in the training of ResNext50-UNet, the trained model still achieves good effect on the segmentation of

group-raised pigs. Specifically, the model after adding the CSA block was able to have better predictive performance for pig individuals far away from the camera (the second and fourth rows in Figure 9). In the case of debris covering the pig's body, both the channel and spatial attention blocks can effectively eliminate the influence of debris on the learning of semantic information in other parts of the pig (the fifth and sixth rows in Figure 9). The above fully proves that the individual pig image segmentation model with the attention block can be effectively transferred to the field of group-raised pigs, which can provide pre-segmentation for group-raised pigs and provide a reference for subsequently refined segmentation.

4 Conclusions

A series of attention-based blocks are proposed for individual pig image segmentation, aiming to adaptively encode rich semantic feature information. Specifically, channel and spatial attention blocks are introduced to capture long-range dependencies from channel and spatial dimensions, respectively. Firstly, experiments on incorporating multiple attention blocks in different encoder and decoder structures show that concurrent channel and spatial attention can capture contextual information more effectively and bring more performance gains. Specifically, using ResNext50 as the encoder and UNet as the decoder, adding channel and spatial attention blocks at the same time can achieve 98.30% and 96.71% on the F1 and IOU metrics, respectively. Compared with the model only adding channel attention block, the two metrics are improved by 0.13% and 0.22%, respectively. In addition, spatial attention is more effective than channel attention. Compared with channel attention, spatial attention improves the F1 and IOU metrics by 0.16% and 0.30%, respectively, when the model is VGG16-LinkNet. Furthermore, adding attention blocks at different layers of the decoder obtains finer semantic information as the depth of the decoding layer increases. More importantly, the individual pig image segmentation model can be transferred to more complex scenarios, which can provide pre-segmentation for group-raised pig scenes.

Acknowledgments

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

[References]

[1] Guo Y Z, Zhu W X, Jiao P P, Ma C H, Yang J J. Multi-object extraction from topview group-housed pig images based on adaptive partitioning and multilevel thresholding segmentation. Biosystems Engineering, 2015; 135: 54–60.

[2] Xu J Y, Zhou S Y, Xu A J, Ye J H, Zhao A Y. Automatic scoring of postures in grouped pigs using depth image and CNN-SVM. Computers and Electronics in Agriculture, 2022; 194: 106746. doi: 10.1016/j.compag.2022.106746.

[3] He K, Wang D, Tong M, Zhu Z J. An improved GrabCut on multiscale features. Pattern Recognition, 2020; 103: 107292. doi: 10.1016/j.patcog.2020.107292.

[4] Ma L, Ji B, Liu H K, Zhu W X, Li W, Zhang T. Differentiating profile based on single pig contour. Transactions of the CSAE, 2013; 29(10): 168–174. (in Chinese)

[5] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015; 521(7553): 436–444.

[6] Hu J, Shen L, Albanie S, Sun G, Wu E H. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020; 42(8): 2011–2023.

[7] Huang G, Liu Z, Van Der Maaten L, Weinberger K Q. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu: IEEE, 2017; pp.2261–2269. doi: 10.1109/CVPR.2017.243.

[8] Li X, Wang W H, Hu X L, Yang J. Selective kernel networks. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.510–519. doi: 10.1109/CVPR.2019.00060.

[9] Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago: IEEE, 2015; pp.1440–1448. doi: 10.1109/ICCV.2015.169.

[10] Lin T Y, Goyal P, Girshick R, He K M, Dollar P. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020; 42(2): 318–327.

[11] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; pp.7263–7271.

[12] Huang Z L, Wang X G, Wei Y C, Huang L C, Shi H, Liu W Y, et al. CCNet: Criss-cross attention for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020; Early access. doi: 10.1109/TPAMI.2020.3007032.

[13] Zhang H, Dana K, Shi J P, Zhang Z Y, Wang X G, Tyagi A, et al. Context encoding for semantic segmentation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018; pp.7151–7160. doi: 10.1109/CVPR.2018.00747.

[14] Bolya D, Zhou C, Xiao F Y, Lee Y J. YOLACT: Real-time instance segmentation. In: 2019 IEEE International Conference on Computer Vision (ICCV), 2019; pp.9156–9165. doi: 10.1109/ICCV.2019.00925.

[15] Wang X L, Zhang R F, Kong T, Li L, Shen C H. SOLOv2: Dynamic and fast instance segmentation. In: 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 2020; Vancouver, pp.17721–17732.

[16] Jensen D B, Pedersen L J. Automatic counting and positioning of slaughter pigs within the pen using a convolutional neural network and video images. Computers and Electronics in Agriculture, 2021; 188: 106296. doi: 10.1016/j.compag.2021.106296.

[17] Huang E D, Mao A X, Gan H M, Ceballos M C, Parsons T D, Xue Y J, et al. Center clustering network improves piglet counting under occlusion. Computers and Electronics in Agriculture, 2021; 189: 106417. doi: 10.1016/j.compag.2021.106417.

[18] 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.

[19] Wang Z Y, Liu T H. Two-stage method based on triplet margin loss for pig face recognition. Computers and Electronics in Agriculture, 2022; 194: 106737. doi: 10.1016/j.compag.2022.106737.

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

[21] Yan H W, Hu Z W, Cui Q L. Study on feature extraction of pig face based on principal component analysis. INMATEH-Agricultural Engineering, 2022; 68(3): 333–340.

[22] Yan H W, Liu Z Y, Cui Q L, Hu Z W, Li Y W. Detection of facial gestures of group pigs based on improved Tiny-YOLO. Transactions of the CSAE, 2019; 35(18): 169–179. (in Chinese)

[23] Chen F E, Liang X M, Chen L H, Liu B Y, Lan Y B. Novel method for real-time detection and tracking of pig body and its different parts. Int J Agric & Biol Eng, 2020; 13(6): 144–149.

[24] Gan H M, Ou M Q, Zhao F Y, Xu C G, Li S M, Chen C X, et al. Automated piglet tracking using a single convolutional neural network. Biosystems Engineering, 2021; 205: 48–63.

[25] Hu Z W, Yang H, Lou T T. Instance detection of group breeding pigs using a pyramid network with dual attention feature. Transactions of the CSAE, 2021; 37(5): 166–174. (in Chinese)

[26] Xiao D Q, Lin S C, Liu Y F, Yang Q M, Wu H L. Group-housed pigs and their body parts detection with Cascade Faster R-CNN. Int J Agric & Biol Eng, 2022; 15(3): 203–209.

[27] Gan H M, Xu C G, Hou W H, Guo J F, Liu K, Xue Y J. Spatiotemporal graph convolutional network for automated detection and analysis of social behaviours among pre-weaning piglets. Biosystems Engineering, 2022;

217: 102–114.

[28] Yan H W, Liu Z Y, Cui Q L, Hu Z W. Multi-target detection based on feature pyramid attention and deep convolution network for pigs. Transactions of the CSAE, 2020; 36(11): 193–202. (in Chinese)

[29] Chen C, Zhu W X, Steibel J, Siegford J, Han J J, Norton T. Recognition of feeding behaviour of pigs and determination of feeding time of each pig by a video-based deep learning method. Computers and Electronics in Agriculture, 2020; 176: 105642. doi: 10.1016/j.compag.2020.105642.

[30] Chen C, Zhu W X, Oczak M, Maschat K, Baumgartner J, Larsen M L V, et al. A computer vision approach for recognition of the engagement of pigs with different enrichment objects. Computers and Electronics in Agriculture, 2020; 175: 105580. doi: 10.1016/j.compag.2020.105580.

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

[32] He H X, Qiao Y L, Li X M, Chen C Y, Zhang X F. Automatic weight measurement of pigs based on 3D images and regression network. Computers and Electronics in Agriculture, 2021; 187: 106299. doi: 10.106/j.compag.2021.106299.

[33] Liu K, Yang H Q, Yang H, Hu Z W, Meng K. Instance segmentation of group-housed pigs based on recurrent residual attention. Journal of South China Agricultural University, 2020; 41(6): 169–178. (in Chinese)

[34] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017; 39(4): 640–651.

[35] Mou L C, Hua Y S, Zhu X X. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.12408–12417. doi: 10.1109/CVPR.2019.01270.

[36] Sun W H, Huang Z P, Liang M, Shao T F, Bi H Z. Cocoon image segmentation method based on fully convolutional networks. In: The Seventh Asia International Symposium on Mechatronics, 2020; 589: 832–843.

[37] Psota E T, Mittek M, Pérez L C, Schmidt T, Mote B. Multi-pig part detection and association with a fully-convolutional network. Sensors, 2019; 19(4): 852. doi: 10.3390/s19040852.

[38] Yang A Q, Huang H S, Zheng C, Zhu X M, Yang X F, Chen P F, et al. High-accuracy image segmentation for lactating sows using a fully convolutional network. Biosystems Engineering, 2018; 176: 36–47.

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

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

[41] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, 2015; pp.234–241.

[42] Yang A Q, Huang H S, Zhu X M, Yang X F, Chen P F, Li S M, et al. Automatic recognition of sow nursing behaviour using deep learning-based segmentation and spatial and temporal features. Biosystems Engineering, 2018; 175: 133–145.

[43] Yang A Q, Huang H S, Yang X F, Li S M, 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.

[44] 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.

[45] He K M, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017; pp.2980–2988. doi: 10.1109/ICCV.2017.322.

[46] Cai Z W, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt lake: IEEE, 2018; pp.6154-6162. doi: 10.1109/CVPR.2018.00644.

[47] Chen K, Pang J M, Wang J Q, Xiong Y, Li X X, Sun S Y, et al. Hybrid task cascade for instance segmentation. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.4969–4978. doi: 10.1109/CVPR.2019.00511.

[48] Tu S Q, Yuan W J, Liang Y, Wang F, Wan H. Automatic detection and segmentation for group-housed pigs based on PigMS R-CNN. Sensors, 2021; 21(9): 3251. doi: 10.3390/s21093251.

[49] Navaneeth B, Singh B, Chellappa R, Davis L S. Soft-NMS--improving object detection with one line of code. In: 2017 IEEE International Conference on Computer Vision (ICCV), 2017; pp.5562–5570. doi: 10.1109/ICCV.2017.593.

[50] Fu J, Liu J, Tian H J, Fang Z W, Lu H Q. Dual attention network for scene segmentation. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019; pp.3146–3154.

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

[52] Woo S, Park J, Lee J Y, Kweon I S. CBAM: Convolutional block attention module. In: The European Conference on Computer Vision (ECCV), 2018; pp.3–19.

[53] Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE Visual Communications and Image Processing (VCIP), Petersburg: IEEE, 2017; pp.1–4.

[54] Kingma D P, Ba J. Adam: A method for stochastic optimization. arXiv, 2014; arXiv: 1412.6980.

[55] Deng J, Dong W, Socher R, Li L J, Li K, Li F F. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami: IEEE, 2009; pp.248-255. doi: 10.1109/CVPR.2009.5206848.

[56] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth International Conference on 3D Vision (3DV), Stanford: IEEE, 2016; pp.565–571. doi: 10.1109/3DV.2016.79.

[57] Sandler M, Howard A, Zhu M L, Zhmoginov A, Chen L C. Mobilenetv2: Inverted residuals and linear bottlenecks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake: IEEE, 2018; pp.4510–4520. doi: 10.1109/CVPR.2018.00474.

[58] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016; pp.770–778. doi: 10.1109/ CVPR.2016.90.

[59] Xie S N, Girshick R, Dollár P, Tu Z W, He K M. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu: IEEE, 2017; pp.5987–5995.

[60] Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu: IEEE, 2017; pp.936–944. doi: 10.1109/CVPR.2017.106.

[61] Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017; pp.2881–2890.

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