agriculture

MDPI

Article

The Detection of Ear Tag Dropout in Breeding Pigs Using a Fused Attention Mechanism in a Complex Environment

Fang Wang, Xueliang Fu *, Weijun Duan, Buyu Wang and Honghui Li

College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohot 010018, China; 2019302100003@emails.imau.edu.cn (F.W.)
* Correspondence: fuxl@imau.edu.cn

Abstract: The utilization of ear tags for identifying breeding pigs is a widely used technique in the field of animal production. Ear tag dropout can lead to the loss of pig identity information, resulting in missing data and ambiguity in production management and genetic breeding data. Therefore, the identification of ear tag dropout is crucial for intelligent breeding in pig farms. In the production environment, promptly detecting breeding pigs with missing ear tags is challenging due to clustering overlap, small tag targets, and uneven sample distributions. This study proposes a method for detecting the dropout of breeding pigs' ear tags in a complex environment by integrating an attention mechanism. Firstly, the approach involves designing a lightweight feature extraction module called IRDSC using depthwise separable convolution and an inverted residual structure; secondly, the SENet channel attention mechanism is integrated for enhancing deep semantic features; and finally, the IRDSC and SENet modules are incorporated into the backbone network of Cascade Mask R-CNN and the loss function is optimized with Focal Loss. The proposed algorithm, Cascade-TagLossDetector, achieves an accuracy of 90.02% in detecting ear tag dropout in breeding pigs, with a detection speed of 25.33 frames per second (fps), representing a 2.95% improvement in accuracy, and a 3.69 fps increase in speed compared to the previous method. The model size is reduced to 443.03 MB, a decrease of 72.90 MB, which enables real-time and accurate dropout detection while minimizing the storage requirements and providing technical support for the intelligent breeding of pigs.

Keywords: ear tag dropout detection; IRDSC; SENet; Cascade Mask R-CNN; Focal Loss

check for updates

Citation: Wang, F.; Fu, X.; Duan, W.; Wang, B.; Li, H. The Detection of Ear Tag Dropout in Breeding Pigs Using a Fused Attention Mechanism in a Complex Environment. *Agriculture* **2024**, *14*, 530. https://doi.org/10.3390/agriculture14040530

Academic Editor: Claudia Arcidiacono

Received: 2 March 2024
Revised: 24 March 2024
Accepted: 25 March 2024
Published: 27 March 2024

1. Introduction

With the rapid development of the pig breeding industry towards precision, intensification, and intelligence, individual identification of breeding pigs is crucial for genetic breeding, feeding management, disease prevention and control, and other aspects of refinement of farming. Large-scale farms primarily use RFID electronic ear tags to identify individual pigs [1]. However, factors such as differences in ear tag quality, scratches from farm facilities, and biting among pigs lead to the ear tag easily falling off. This causes production and genetic breeding data to be lost or mixed up, affecting intelligent farming management [2]. Therefore, real-time and accurate detection of ear tag dropout is of great significance for the breeding management and genetic breeding in the field of breeding pig production.

With the rapid development of computer vision technology and deep learning [3,4], deep convolutional neural networks show a strong feature extraction ability and are widely used for individual recognition in smart farming [5]. Deep-learning-based target detection networks have been applied to individual recognition [6–8], posture detection [9,10], target tracking [11,12], and counting statistics [13,14] in the field of animal husbandry, achieving better results and verifying the feasibility of deep convolutional neural networks for animal individual recognition. LI Jianquan et al. utilized an improved YOLOv5s model to solve the identification problem of group-raised pigs in real environments, achieving

dual improvements in detection accuracy and speed and offering ideas for individual pig identification in complex environments [15]. However, the common characteristic of these studies is the insufficient consideration of time performance. In summary, existing studies are mainly devoted to the accuracy of detection and identification, and the effectiveness in complex environments still needs to be further improved. No relevant research has been conducted on ear tag dropout detection in breeding pigs.

This study presents a novel method for identifying ear tag dropout in breeding pigs in complex environments. This approach reduces model parameters and enhances feature extraction capabilities through depthwise separable convolution [16] and an inverted residual structure [17]. A channel attention mechanism [18] is integrated for deep semantic feature enhancement. Finally, an ear tag dropout detection model based on Cascade Mask R-CNN [19] is designed to improve the accuracy and detection speed.

2. Materials and Methods

2.1. Data Acquisition

Images of breeding pigs were captured in December 2022 at a large-scale breeding pig farm in Hohhot, Inner Mongolia Autonomous Region, China. The pigsty accommodated 28 breeding pigs aged 2 to 3 months; 2 pigs had missing ear tags, while the remaining 26 pigs had intact ear tags. A fixed-position image collection system was set up for the experiment utilizing a hemispherical camera (DS-2PT7D20IW-DE, Hikvision, Hangzhou, China) mounted 3.4 m above the feeding area in the pigsty (refer to Figure 1). The camera boasted a resolution of 1920 pixels by 1080 pixels and captured images of breeding pigs in their natural environment. The video footage was stored in the farm's local area network, NVR, in ASF format and encoded with H.264.

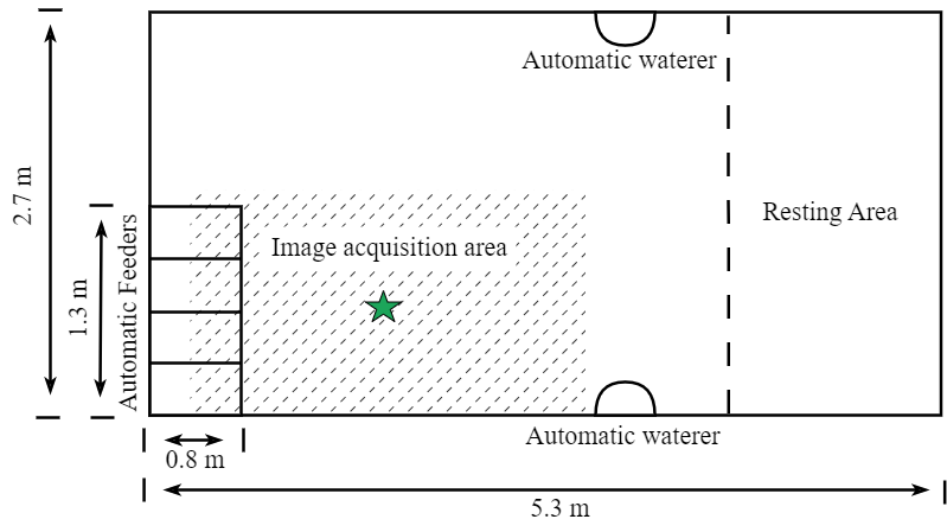


Figure 1. The layout of the pig pens during the experiment is illustrated in the diagram. The green star represents the location of the camera. The shaded area represents the image acquisition area, which is also the feeding area for the breeding pigs; all of the breeding pigs pass through this area when feeding.

2.2. Establishment of the Data Sample Database

2.2.1. Data Filtering

Sixty-four video segments featuring breeding pigs with ear tag dropout were manually selected from the collected video files. Using FFMPEG software (3.2.19), one frame was extracted per second, resulting in 96,087 images. To enhance the model training efficiency, the images were cropped to focus on the feeding area of the pigs, resulting in a cropped images size of 1438 × 973 pixels. Due to the high similarity between adjacent frames, utilizing them directly for model training could lead to overfitting. Therefore, this study

utilized the structural similarity (SSIM) [20] algorithm to filter out highly similar images, as demonstrated in Equation (1).

$$SSIM(f_1, f_2) = \frac{(2\mu_1\mu_2 + c_1)(2\sigma_{1,2} + c_2)}{(\mu_1^2 + \mu_2^2 + c_1)(\sigma_1^2\sigma_2^2 + c_2)}, \quad (1)$$

where μ_1 and μ_2 represent the pixel averages of images f_1 and f_2, respectively; σ_1^2, σ_2^2 denote the pixel variance of images f_1, f_2; $\sigma_{1,2}$ is the pixel covariance between images f_1 and f_2; and c_1 and c_2 are two constants used to avoid division by zero, with their values both set to 0.01. The experiment established an empirical threshold for SSIM at 0.78, resulting in the filtering of 92% of highly similar images. Subsequent manual selection was carried out to eliminate images that were either damaged or did not feature breeding pigs, resulting in the retention of 5865 images of breeding pigs. The dataset was then split into a training set (4692 images) and a test set (1173 images) at a ratio of 4:1.

2.2.2. Image Annotation

The breeding pigs were masked and annotated using the open-source image annotation software PaddleSeg (V2.7), with the results saved in the COCO dataset [21] format. The training set consisted of 10,894 annotated breeding pigs, with 4605 with ear tags and 6289 without. There were 2671 annotated breeding pigs in the test set, including 1121 with ear tags and 1550 without. Figure 2 illustrates the distribution of the data.

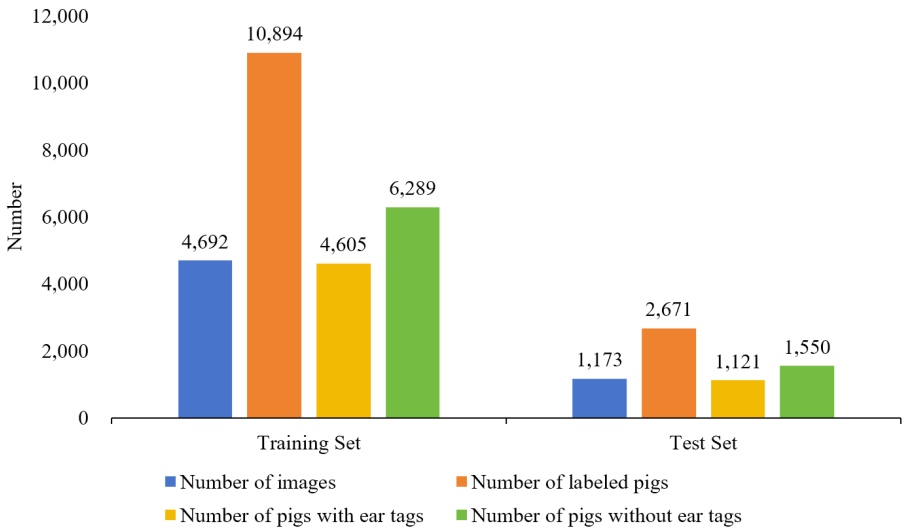


Figure 2. Distribution of the number of experimental datasets.

2.2.3. Data Enhancement

To verify the robustness of the model in complex environments, the Python data augmentation library albumentations [22] and single-sample data augmentation methods were used to augment the images in the test set equally. This includes vertical flips, horizontal flips, Gaussian blur (convolution kernel with a size of 5 × 5 and standard deviation of 3.0), contrast adjustment (with adjustment factors randomly sampled within the range of −50 to 100), modification of pixel values (randomly altering RGB channel values within the range of −50 to 50), and random occlusion (covering with a black square whose side length ranged from 10 to 300 pixels). This generated six groups of comparative test data, as shown in Figure 3.

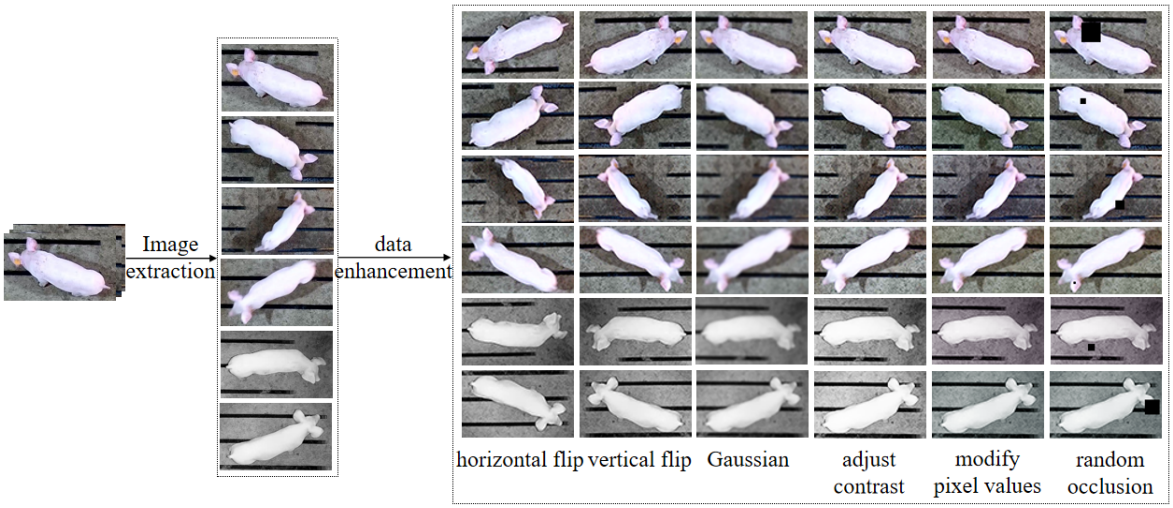


Figure 3. Image enhancement examples.

2.3. Design of the Model

Cascade Mask R-CNN enhances the performance of object detection and instance segmentation by incorporating a cascade structure based on Mask R-CNN [23]. Our research team has proposed an improved Cascade Mask R-CNN algorithm, which showed better results in detecting ear tag dropout for breeding pigs in a production environment, verifying the feasibility of conducting related research based on Cascade Mask R-CNN. However, this model also faces challenges such as a large scale, leading to longer training and inference times that require more computational resources. Enhancing the model's detection accuracy and processing speed enables instant analysis of video streams or numerous images, facilitating real-time and precise ear tag dropout detection and aiding farms in prompt intervention measures, which is crucial for enhancing farm management efficiency and animal welfare. Therefore, this paper proposes a novel ear tag dropout detection model for breeding pigs called Cascade-TagLossDetector, based on Cascade Mask R-CNN. This model optimizes the detection accuracy and speed by integrating a lightweight feature extraction module (IRDSC) and a feature enhancement module (SENet). It consists of three components: a backbone network for semantic feature extraction, a region suggestion network for generating candidate target region bounding boxes, and a cascade detection network for cascading detection and correction of candidate target regions.

2.3.1. The Structure of Cascade-TagLossDetector

Cascade-TagLossDetector utilizes ResNeXt101 and a feature pyramid network as the backbone network for feature extraction, improving the model's capability of multi-scale feature extraction in intricate pigsty environments. ResNeXt101, with its deeper network structure and more intricate topology compared to ResNet50, offers improved learning and extraction abilities for complex and abstract features. It is composed of 4 groups of residue blocks in a series, with each group containing 3, 4, 23, and 3 residue blocks, respectively. Within the residual blocks, three convolution operations of 1×1, 3×3, and 1×1 are utilized, with the second 3×3 convolution employing channel group convolution with 32 branches and 64 groups. Previous research has shown that depthwise separable convolution can reduce the model parameters and enhance feature extraction capabilities, while the inverted residual structure excels in capturing features at various scales. This study combines depthwise separable convolution and an inverted residual structure to create a lightweight feature extraction module, IRDSC, which replaces standard convolution in ResNeXt101. This aims to improve the feature extraction efficiency and decrease the computational load. Furthermore, the channel attention mechanism SENet is employed within each convolutional layer of the grouped convolution structure to capture relationships between different channels, improving the model's representation of crucial

features and enabling adaptive learning of channel importance, thereby enhancing the feature expression capability.

The region suggestion network generates anchor frames for candidate target regions based on the outputs of the five 256-channel feature layers of the FPN network. It then performs target classification and coordinates regression by mapping these anchor frames back to the original image space.

The cascade detection network trains detectors in a cascaded manner to enhance the model's ability to recognize and localize targets. This study focuses on the challenge of the limited number of pigs with missing ear tags and the uneven distribution of positive and negative samples in the training data, which hinders detection accuracy improvements, and utilizes the advantages of the Focal Loss classification accuracy for specific categories. This study optimizes the classification and regression loss calculation approach for cascade detection, enhancing the model's focus on detecting rare and more challenging samples. Additionally, dropout [24] is applied in the cascaded detection head by randomly discarding neurons to reduce network overfitting. A fully connected layer performs weighted summation and linear transformation of neurons in the previous layer for target detection and regression prediction. The structure of Cascade-TagLossDetector is illustrated in Figure 4.

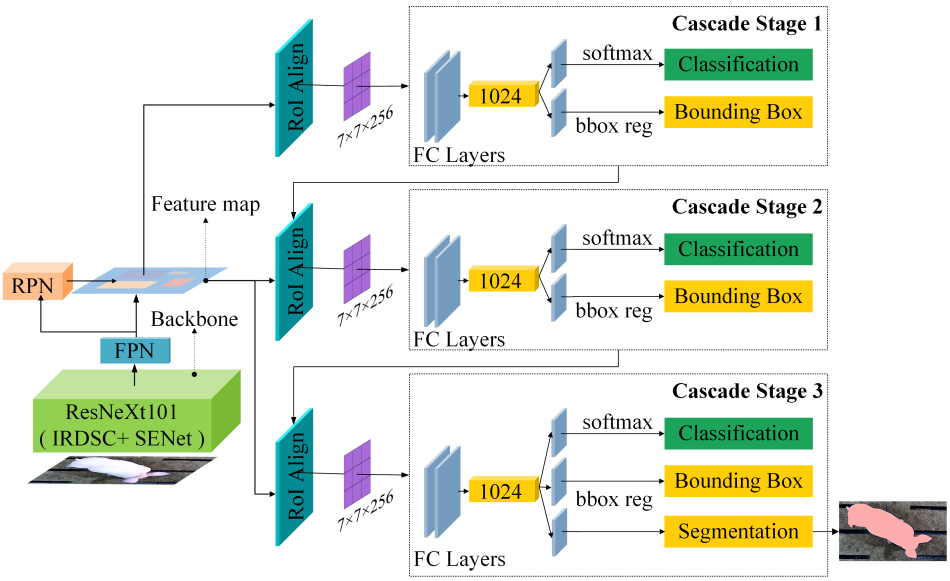


Figure 4. The model structure of Cascade-TagLossDetector.

2.3.2. Lightweight Feature Extraction Module IRDSC

The experimental image contains multi-layer semantic information; extracting rich semantic information from the image is crucial for real-time and accurate detection of ear tag dropout. Utilizing ResNeXt101 and feature pyramid networks as the feature extraction backbone network can improve the network's expressive power and accuracy. However, this approach also leads to an increased computational complexity and parameter count. Therefore, this study focuses on leveraging depthwise separable convolution and an inverted residual structure to develop a lightweight feature extraction module called IRDSC (see Figure 5). Depthwise separable convolution captures spatial correlations within the input channels by performing convolution operations on each input channel through deep convolution. Pointwise convolution then applies a 1×1 convolution kernel on the output of the deep convolution to capture correlations and feature combinations across channels.

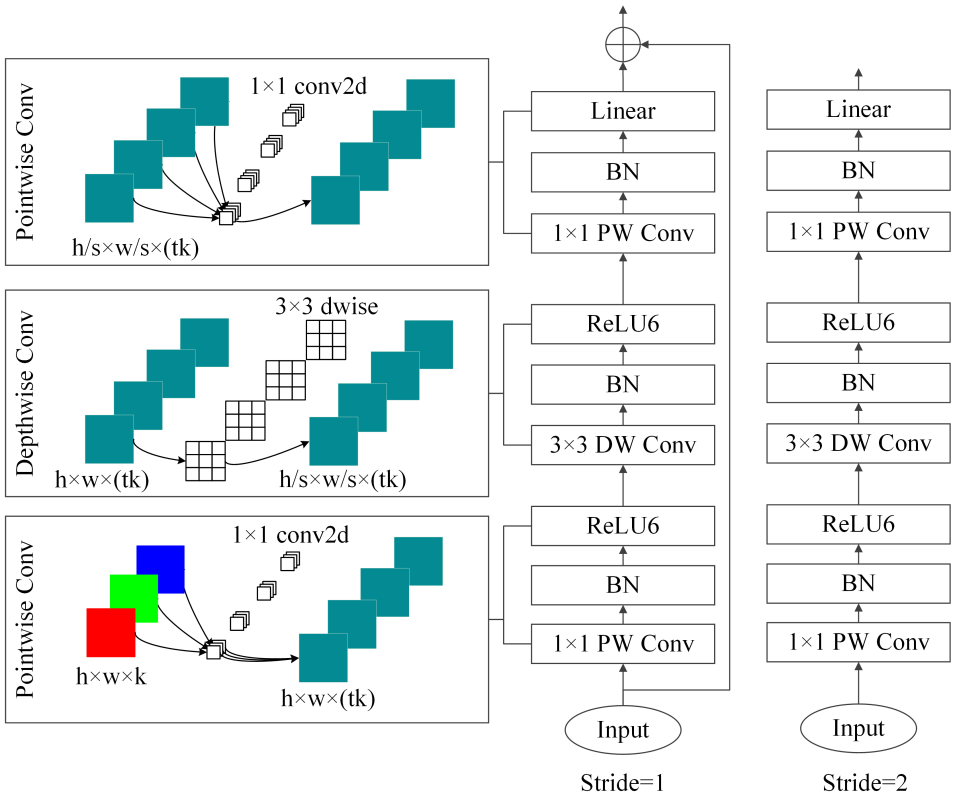


Figure 5. Structure diagram of the IRDSC.

The IRDSC module initiates linear activation and batch normalization operations to standardize the distribution of feature channels, reducing internal covariate shifts and expediting training convergence. Following this, it expands the number of feature map ($h \times w \times k$) channels to tk through 1×1 pointwise convolution and then applies a 3×3 deep convolution with a stride of s to extract features from each channel. Subsequently, the feature map channels are reduced to k' using 1×1 pointwise convolution to match the input and output channel dimensions to reduce the model parameters and computational complexity. Additionally, the module establishes residual connections between input and output feature maps to preserve original information and enhance feature extraction capabilities. The algorithm employs a linear activation function ReLU6 [25] to prevent information loss during dimensionality reduction from high-dimensional to low-dimensional spaces, as illustrated in Equation (2).

$$f(x) = \min(\max(0, x), 6) \tag{2}$$

ReLU6 constrains feature values within a range of 0 to 6. Values exceeding 6 are limited to 6 to avoid gradient explosion, while values below 0 are adjusted to 0 to prevent gradient disappearance. This restriction helps stabilize the model's training process, enhance the model's learning abilities, and improve the overall robustness.

2.3.3. Semantic Feature Enhancement Module SENet

The enhanced lightweight feature extraction module improves the feature extraction capability while reducing model parameters, but there is also a certain feature loss. Therefore, this study integrates the channel attention mechanism SENet for deep semantic feature enhancement. SENet maintains attention generation and feature enhancement by adaptively boosting useful features and suppressing irrelevant ones. This lightweight module helps prevent parameter count proliferation. The structure of SENet is illustrated in Figure 6.

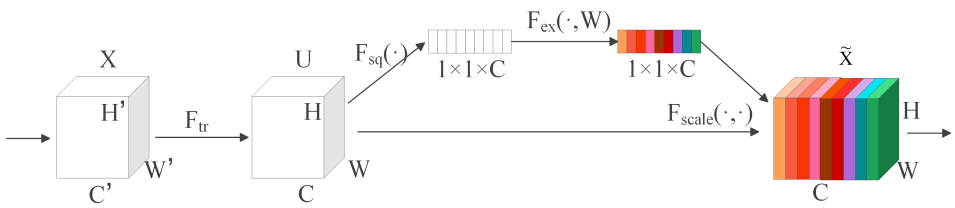


Figure 6. Structure diagram of SENet.

(1) The feature map X ($W' \times H' \times C'$) extracted by the backbone network is input into the SENet module. A new feature map U ($W \times H \times C$) is generated through a 1×1 depthwise separable convolution, calculating the spatial correlation of the channels with a set of filter parameters V, as shown in Equation (3).

$$U_C = V_C * X = \sum_{S=1}^{C'} V_C^S * X^S \tag{3}$$

where X denotes the input feature map, U denotes the output feature map, C is the number of channels, W and H denote the width and height of the feature map, $V = [V_1, V_2, \ldots, V_c]$ denotes the set of learned filters, V_C stands for the Cth filter, * is a convolution operation, V_c^s is the 2D spatial kernel acting on the corresponding channel of X, and X^S denotes the Sth input. After the transformation, U is the matrix that contains the C $W \times H$ feature maps and U_C is the Cth feature map in the matrix.

(2) Squeeze conducts spatial dimensional feature compression on the feature map through global average pooling. This process converts the 2D feature channels into a single real number with a global receptive field, effectively compressing the $W \times H \times C$ feature map into a $1 \times 1 \times C$ feature vector, as shown in Equation (4).

$$Z_C = F_{sq}(U_C) = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} U_C(i,j) \tag{4}$$

where Z_C denotes the local description operator of the compressed feature information distribution.

(3) Excitation utilizes the fully connected layer operation W_1 to decrease the number of channels, which is activated by ReLU, then passes through the fully connected layer W_2 to restore the number of channels, and finally applies the Sigmoid function to ensure that the channel attention weight s takes a value in the range of [0, 1], outputting the weight coefficients that indicate the importance of the channel.

$$S = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)) \tag{5}$$

In Equation (5), W_1 denotes the $R^{C/r*C}$ weight matrix, W_2 is the $R^{C*C/r}$ weight matrix, r is the dimensionality reduction parameter, which is set to 16 experimentally, δ is the ReLU activation function, and σ denotes the Sigmoid function. W_1 reduces the dimensionality of $1 \times 1 \times C$ to $1 \times 1 \times C/r$, and W_2 restores it back to $1 \times 1 \times C$. The number of parameters and the amount of calculations of the algorithm are effectively reduced after the operation of the two fully connected layers.

(4) Scale multiplies the channel attention weights with the original feature map channel by channel for feature recalibration; this process ensures that important information receives attention closer to 1, while less relevant information is closer to 0. By obtaining a feature map with attention weights, the model can dynamically enhance the weights of important channels and suppress the weights of non-important channels to achieve the effect of feature enhancement.

2.3.4. The Loss of the Model

Cascade Mask R-CNN utilizes binary cross-entropy loss for classification and regression losses in the region proposal network. The cascaded detection network employs binary cross-entropy loss, smooth L1 loss, and pixel-level binary cross-entropy loss for classification, regression, and mask losses. This study integrates the Focal Loss function in the cascaded detection heads for calculating classification and regression losses. Focal Loss introduces an adjustable balancing parameter and a factor to modulate weights for difficult and easy samples, enabling high-precision detection of common and rare samples, as depicted in Equation (6).

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{6}$$

This equation includes various parameters such as p_t for the predicted probability, α_t for the equilibrium parameter, and γ as the moderator. The term $(1-p_t)^{\gamma}$ is responsible for suppressing the loss of easy-to-categorize samples in the model and focusing more on hard-to-categorize samples.

Based on this, the loss of the Cascade-TagLossDetector model comprises the classification loss L_{rpn_cls} and regression loss L_{rpn_reg} of the region proposal network, as well as the classification loss $L_{\_cls}$, regression loss $L_{\_reg}$, and mask loss $L_{\_mask}$ of the cascaded detection. The model's total loss is the weighted sum of these component losses, as illustrated in Equation (7).

$$L_{cascade} = L_{rpn\_cls} + L_{rpn\_reg} + \sum_{i=1}^{n} \lambda_i ({L_{\_cls}}^i + {L_{\_reg}}^i) + L_{\_mask} \tag{7}$$

The weights λ_i for each cascade are set to 1, 0.5, and 0.25, respectively.

2.3.5. Evaluation Indicators

This study evaluates the recognition performance of the model through metrics such as Precision (P), Recall (R), Average Precision (AP), bounding box detection mean average precision (bbox mAP), and instance segmentation mean average precision (mask mAP), and the computational performance of the model is also comprehensively evaluated using the detection speed (Speed), the inference time (Inference time), and the storage space occupied by the model (Model_Size), see Equations (8)–(11). Furthermore, the accuracy of the model in detecting ear tag dropout in breeding pigs is evaluated by defining the accuracy as the ratio of correctly recognized breeding pigs with ear tag dropouts to breeding pigs without ear tag dropouts, as shown in Equation (12).

$$precision = \frac{TP}{TP+FP} \tag{8}$$

$$recall = \frac{TP}{TP+FN} \tag{9}$$

$$AP = \int_0^1 P(R)dR \tag{10}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N}\left(\int_0^1 p_i(r)dr\right) \tag{11}$$

$$Accuracy = \frac{TP_{ET\_Drop} + TP_{ET\_NotDrop}}{N_{ET\_Drop} + N_{ET\_NotDrop}} \tag{12}$$

AP is the area of the region enclosed by the precision–recall curve and the x-axis and y-axis; mAP is calculated as the mean value of AP with an IoU threshold ranging from 0.50 to 0.95 in steps of 0.05. In the experiment, pigs with ear tags are considered positive samples, while those with missing ear tags are considered negative samples. TP, FP, and FN

represent the categories of true positives, false positives, and false negatives, respectively. In Equation (12), N_{ET_Drop} refers to the number of breeding pigs with missing ear tags, $N_{ET\_NotDrop}$ represents the number of breeding pigs with present ear tags, TP_{ET_Drop} indicates the number of breeding pigs with missing ear tags that were correctly identified, and $TP_{ET\_NotDrop}$ signifies the number of breeding pigs with present ear tags that were correctly identified.

2.3.6. Design of Experiments

The experiment utilized two Intel(R) Xeon(R) Gold 6137 processors with 256 GB of RAM and eight NVIDIA GeForce RTX 3090 graphics cards for training. The system ran on Ubuntu 20.04 OS with Miniconda3, Python 3.8.5, CUDA 11.7, Pytorch 2.0.0, and MMDetection 2.28.2 to establish a deep learning algorithm framework. Stochastic gradient descent was used as the optimizer with an initial learning rate of 0.02, a momentum coefficient of 0.9, and a weight decay coefficient of 0.0001 for a total of 300 training epochs and a batch size of 64.

In order to validate the effectiveness of the model for ear tag dropout detection in breeding pigs, the experiment initially involved training Cascade Mask R-CNN and Cascade-TagLossDetector on the training set. The training process utilized stochastic gradient descent to adjust the parameters and determine the optimal model. Subsequently, both the original test data and the six sets of enhanced test data were fed into the model for detection, as shown in Figure 7.

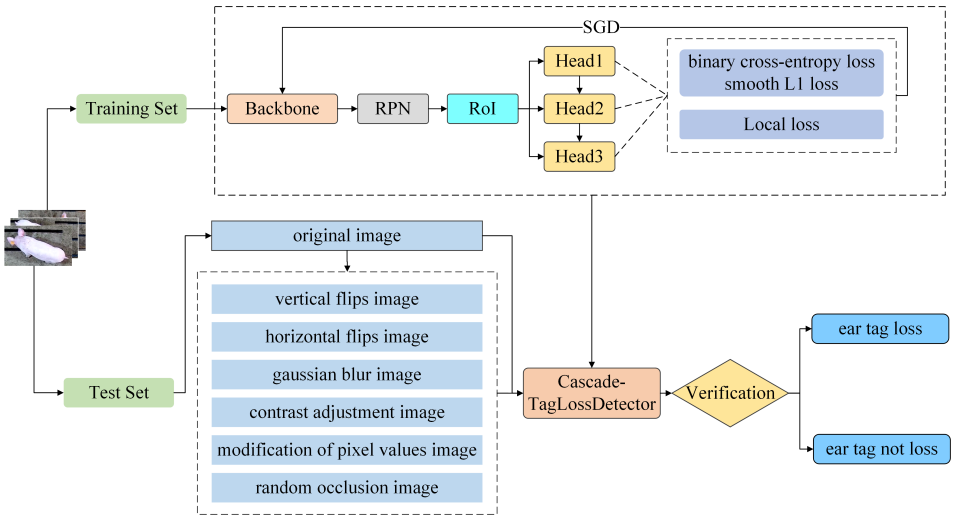


Figure 7. Cascade-TagLossDetector's process of ear tag dropout detection in breeding pigs.

(1) In this study, a labeled individual positioned at a minimum distance of 1 px from the image edge is defined as a breeding pig fully within the detection field. The experiment specifically targeted the detection of ear tag dropout in breeding pigs within the field of view.

(2) N_{ET_Drop}, $N_{ET\_NotDrop}$, TP_{ET_Drop}, and $TP_{ET\_NotDrop}$ of the model were computed on the original test set, respectively, to calculate the accuracy.

3. Results

3.1. The Results of Training

To assess the effectiveness of the proposed model, this experiment evaluates the training results of both Cascade Mask R-CNN and Cascade-TagLossDetector and compares the bbox mAP, mask mAP, and loss values; the findings are depicted in Figure 8.

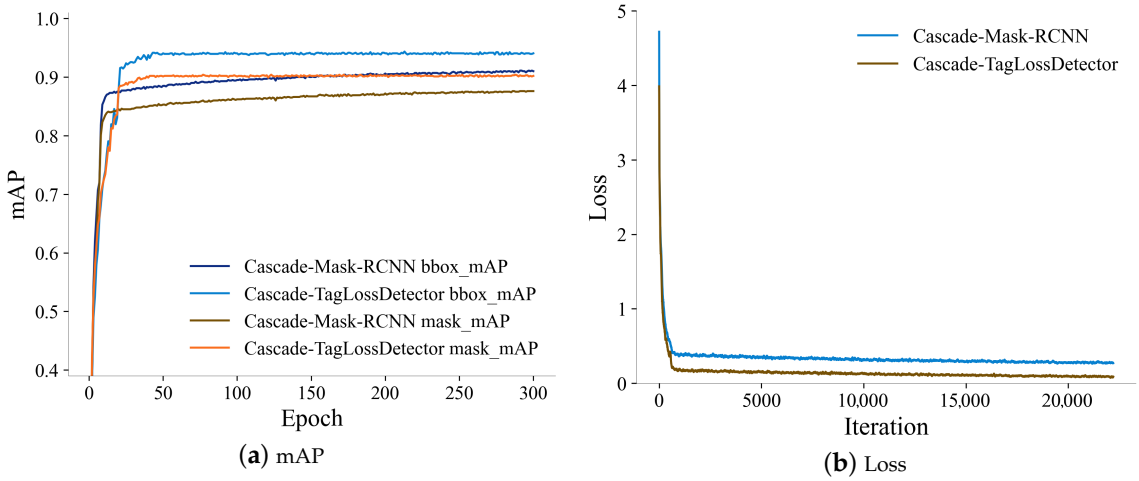


Figure 8. The parameter change curves during training.

Figure 8 illustrates that both models show an upward trend in bbox mAP and mask mAP as the number of training epochs increases, accompanied by a decrease in loss values that eventually stabilizes. The bbox mAP and mask mAP of the model proposed in this study stabilize at around 94% and 90% respectively, after 50 epochs, demonstrating a notable improvement over Cascade Mask R-CNN. The loss values of Cascade-TagLossDetector settle around 0.1 after 1500 iterations, marking a reduction of approximately 0.2 compared to Cascade Mask R-CNN's loss values, thus validating the efficacy of the proposed model.

3.2. Performance Analysis of the Improvement Strategy

This paper enhances the Cascade Mask R-CNN model by incorporating ResNeXt101 as the feature extraction backbone network, along with IRDSC and SENet. Focal Loss is utilized to compute classification and regression losses in cascade detection, replacing the original loss functions. Ablation experiments were conducted to determine the optimal model for recognizing ear tag dropout in breeding pigs, with the results compared in Table 1.

Table 1. Results of ablation experiments .

Model	Loss Function	bbox mAP/%	mask mAP/%	Model Size/MB	Recall/ %	Accuracy/ %	Inference Time/ms	Speed/ (f/s)
Cascade Mask R-CNN	Original loss	91.10	87.14	515.93	91.14	87.07	46.21	21.64
	Focal Loss	91.14	87.63	515.93	91.68	87.45	46.21	21.64
Cascade Mask R-CNN + IRDSC	Original loss	93.19	88.02	414.05	93.79	88.34	36.74	27.22
	Focal Loss	93.24	88.59	414.05	94.67	88.98	36.74	27.22
Cascade Mask R-CNN + IRDSC + SENet	Original loss	94.01	88.37	443.03	94.88	88.92	39.48	25.33
	Focal Loss	94.15	90.32	443.03	97.42	90.02	39.48	25.33

Analysis of Table 1 demonstrates that incorporating Focal Loss for classification and regression losses in cascade detection can improve the mean average precision of ear tag dropout recognition in breeding pigs compared to the original loss functions without impacting the model size or computational performance. This highlights Focal Loss's effectiveness in addressing class imbalance issues. Integrating the lightweight feature extraction module IRDSC into the Cascade Mask R-CNN backbone network and combining it with the Focal Loss function achieved a bbox mAP and mask mAP of 93.24% and 88.59%, respectively, marking improvements of 2.14 and 1.45 percentage points. The model size was reduced by 101.88 MB, and the computation speed increased by 5.58 frames per second. Substituting standard convolutions with depthwise separable convolutions and inverted residual structures enhanced the model's ability to extract multi-scale information

features and optimizes the computational performance. The addition of the semantic feature enhancement module SENet allowed for adaptive amplification of useful features and suppression of irrelevant ones, resulting in bbox mAP and mask mAP values of 94.15% and 90.32%, respectively, with the model's accuracy reaching an impressive 90.02%. The model size was reduced to 443.03 MB, a decrease of 72.90 MB compared to Cascade Mask R-CNN. The IRDSC and SENet modules significantly enhance the model's accuracy and computational performance in recognizing ear tag dropout in breeding pigs.

3.3. The Comparison Experiment of the Test Set

To assess the model's robustness under various environmental conditions, experiments were conducted to verify the model's performance on the original test set, and datasets were altered through vertical flipping, horizontal flipping, Gaussian blurring, contrast adjustment, pixel value modification, and random occlusion. As depicted in Table 2, the models exhibited the highest detection accuracy on the original dataset; the bbox mAP and mask mAP of the proposed model were 94.17% and 90.26%, respectively, closely resembling the performance on the training set. This suggests that the model demonstrates good generalization without signs of overfitting or underfitting. The accuracy of the model for ear tag dropout detection in breeding pigs is 90.02%, which is 2.95% higher than that of Cascade Mask R-CNN. Operations such as vertical flipping, horizontal flipping, pixel value modification, and random occlusion have minimal impacts on the deep network, with the presented model achieving bbox mAP values of 93.84%, 94.01%, 75.12%, and 68.59%, and mask mAP values of 90.21%, 90.13%, 65.14%, and 61.20%, respectively, which are all significantly higher than those of Cascade Mask R-CNN. On datasets with Gaussian blurring and an adjusted contrast, both models exhibited a decrease in bbox mAP and mask mAP, with the proposed model showing a lower decrease than Cascade Mask R-CNN. The feasibility of using Cascade-TagLossDetector for ear tag dropout recognition in breeding pigs in complex environments was also confirmed.

Table 2. Detection results of the seven test sets.

Data Set	Cascade Mask R-CNN			Cascade-TagLossDetector		
	bbox mAP/%	mask mAP/%	Accuracy/%	bbox mAP/%	Mask mAP/%	Accuracy/%
Original test set	91.10	87.14	87.07	94.17	90.26	90.02
Vertical flip dataset	90.79	86.54	85.19	93.84	90.21	89.96
Horizontal flip dataset	91.08	87.12	86.34	94.01	90.13	89.97
Gaussian blur dataset	26.02	33.54	53.64	30.32	38.89	57.88
Contrast-adjusted dataset	34.59	40.88	62.70	43.68	46.21	64.29
Pixel-value- modified dataset	71.09	64.32	74.21	75.12	65.14	77.01
Randomly occluded dataset	62.33	48.65	68.34	68.59	61.20	73.64

4. Discussion

In this experiment, the hierarchical 10-fold cross-validation method [26] is utilized to assess the model's performance in identifying ear tag dropout in breeding pigs to avoid detection and evaluation errors caused by data imbalances. The experiment involved dividing seven test sets into ten equal parts, creating ten datasets with a mix of seven types of data by selecting one part from each set; nine datasets were utilized for training, while the tenth dataset was used for testing, focusing on bbox mAP, mask mAP, and accuracy. This process was iterated ten times, with average values calculated to determine the final bbox mAP, mask mAP, and accuracy on the test set. The findings are summarized in Table 3.

Table 3. The results of layered 10-fold cross-validation.

Test	Cascade Mask R-CNN			Cascade-TagLossDetector		
	bbox mAP/%	mask mAP/%	Accuracy/%	bbox mAP/%	Mask mAP/%	Accuracy/%
The first fold	86.8818	82.5505	82.6103	91.1250	88.1365	88.7104
The second fold	91.2304	87.1494	87.4378	93.9169	90.8662	89.9653
The third fold	90.7221	86.9565	86.7934	93.0694	89.0221	89.0211
The fourth fold	89.9554	86.1201	86.0121	93.2161	87.9151	87.6392
The fifth fold	88.2361	84.3695	84.5422	90.1264	86.9664	86.4849
The sixth fold	93.9949	89.9856	89.8145	96.4694	93.3101	92.3201
The seventh fold	91.1316	88.0221	88.1306	95.0661	90.8685	90.8741
The eighth fold	92.6651	89.0754	89.4201	96.6394	91.3155	91.5302
The ninth fold	95.0613	89.5404	88.4509	96.5664	92.8681	92.1042
The tenth fold	91.1163	87.6513	87.4922	95.2909	91.9105	91.5072
Average value	91.0995	87.1421	87.0704	94.1486	90.3179	90.0157
Variance	6.00	5.38	4.88	5.22	4.78	3.97

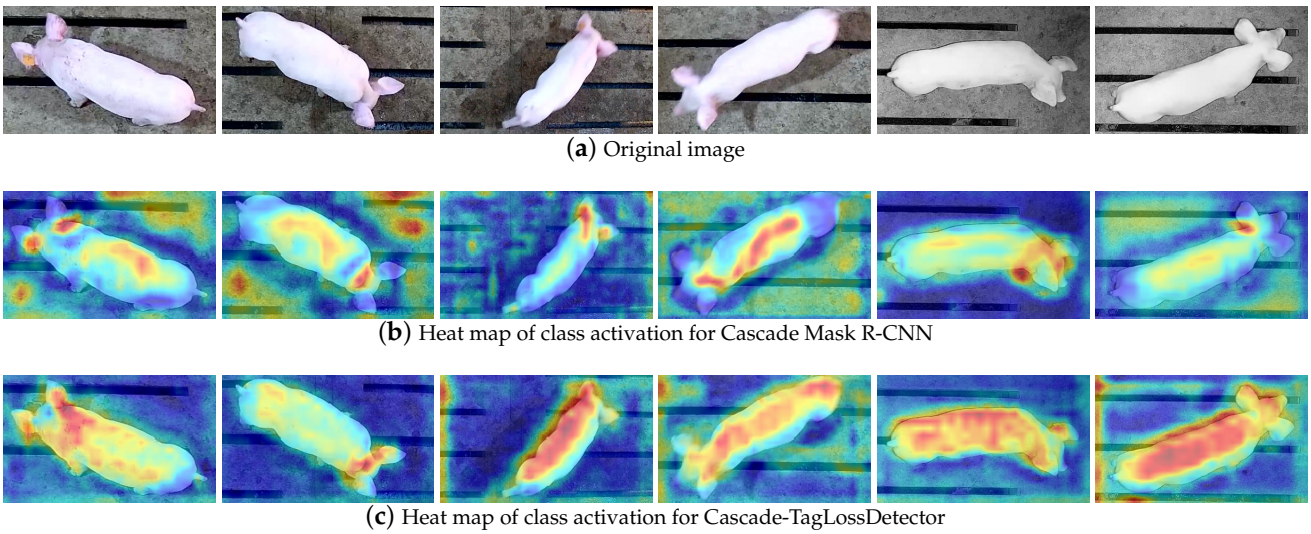
Analysis of Table 3 reveals that the bbox mAP and mask mAP of the two models remained consistent within a certain range throughout the hierarchical ten-fold cross-validation process, mainly because each dataset contains the same number of original and data-enhanced images, whereas Gaussian blurring, the adjustment of contrast, modifications of pixel values, and random occlusion randomly shift the values to within a certain range for data enhancement. Differences between the data lead to fluctuations up and down in the cross-validation results of the same evaluation indexes. Notably, Cascade-TagLossDetector demonstrates a superior detection segmentation performance in each validation, mainly due to SENet's ability to retain detailed features of the deep semantics, and this conclusion is consistent with Lu Zhou et al.'s findings [27]. Furthermore, our team's earlier study suggests [28] that the enhanced detection segmentation performance is linked to the utilization of ResNeXt101 as the backbone network, compared to Cascade Mask R-CNN which utilizes ResNet50. ResNeXt101 offers a deeper network level, facilitating the extraction of multi-scale feature information, and the 32-branching strategy in the second convolution of ResNeXt101's residual block effectively prevents the loss of detailed information, enhancing detection effectiveness. Additionally, the model's bbox mAP, mask mAP, and accuracy exhibit a lower variance (5.22, 4.78, and 3.97, respectively) compared to Cascade Mask R-CNN, indicating that the model is more stable.

By analyzing Tables 1–3, it can be found that there is a positive correlation between model accuracy and the mean average precision (mAP). For instance, when the model was upgraded from Cascade Mask R-CNN to Cascade-TagLossDetector, there was a 3.18% increase in mask mAP, accompanied by a 2.95% rise in model accuracy. This phenomenon reveals an insight: as an indicator that comprehensively reflects the model's detection performance, the improvement in mAP directly contributes to an increase in the model's overall accuracy on the test set.

The recognition results of the model on both the training dataset and the test datasets in Table 2 are similar; this similarity can be attributed to the use of the IRDSC module, which incorporates a linear activation function to mitigate gradient vanishing. Additionally, the utilization of fully connected linear transformations, dropout layers, and ReLU activation functions helps to prevent overfitting.

Xie Qiuju et al. illustrated that the introduction of an attention module could strengthen the model's attention to the features of the pig's face through class-activated heat maps [29]. Thus, the experiment was visualized using a class-activated heat map [30] on the feature extraction of the ear tag dropout in breeding pigs, where warmer-colored regions indicate higher attention from the model. Figure 9 shows breeding pigs' head and spine parts, particularly the head and ear regions, highlighted in warm colors. This suggests that the inclusion of IRDSC, SENet, and the Focal Loss loss function prompts the model to pay closer attention to fine-grained original data features during feature abstraction, leading

to a higher activation of important features in these areas and ultimately improving the detection accuracy.



(**a**) Original image

(**b**) Heat map of class activation for Cascade Mask R-CNN

(**c**) Heat map of class activation for Cascade-TagLossDetector

Figure 9. Heat map of the model.

The model proposed in this study, based on Cascade Mask R-CNN, has been improved regarding feature extraction and enhancement, which improves the accuracy of ear tag dropout detection in pigs and overcomes environmental interference to some extent, making it valuable for pig breeding and genetic management. However, the model did not completely solve the problems of misdetection and omission and did not realize the identification and tracking of breeding pigs with missing ear tags. Future research will concentrate on further refining the network model to enhance the accuracy and minimize the model size, as well as develop methods for identifying and tracking pigs with missing ear tags.

5. Conclusions

This paper presents a method for detecting ear tag dropout in breeding pigs in complex environments by integrating attention mechanisms. The algorithm is based on Cascade Mask R-CNN and incorporates IRDSC and SENet into the backbone network, optimizing the loss function with Focal Loss. The model achieves a bbox mAP of 94.15% and a mask mAP value of 90.32%, outperforming Cascade Mask R-CNN by 3.05% and 3.18%, respectively. The detection accuracy is 90.02%, which is also improved by 2.95%, with an increase in the detection speed of 3.69 fps and a reduction in the model size of 72.90 MB. Experimental results demonstrate that enhancing the feature extraction capability of the backbone network with IRDSC and reducing the model size is feasible; SENet and Focal Loss can realize the enhancement of deep semantic information and key features. The proposed model performs better than Cascade Mask R-CNN on six enhanced test sets, confirming its effectiveness in detecting and recognizing ear tag dropout in pigs in complex environments and providing valuable insights for intelligent breeding in pig farms.

Author Contributions: F.W.: conceptualization, funding acquisition, investigation, methodology, software, writing—original draft, writing—review and editing. X.F.: funding acquisition, resources, supervision. W.D.: data curation, formal analysis, investigation, software, validation. B.W.: data curation, formal analysis, investigation. H.L.: conceptualization, funding acquisition, methodology. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Key Science and Technology Special Project of Inner Mongolia Autonomous Region (2021ZD0005), the Special Project for Building a Science and Technology Innovation Team at Universities of Inner Mongolia Autonomous Region (BR231302), the National Natural Science Foundation of China (61962047), and the Research Innovation Foundation of Graduate Students of Inner Mongolia Autonomous Region (BZ2020054).

Institutional Review Board Statement: Ethical review and approval, as well as Institutional Review Board (IRB) statement, are not applicable for this study. This research is based on visual detection of ear tag dropout in breeding pigs and does not involves direct contact with the animals or any intervention that could potentially harm them. The study solely utilizes visual data for the purpose of improving the accuracy of ear tag detection, thus does not fall under the scope of ethical review that pertains to studies involving live subjects.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available from the corresponding author on reasonable request.

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

References

1. Adrion, F.; Kapun, A.; Eckert, F.; Holland, E.M.; Staiger, M.; Götz, S.; Gallmann, E. Monitoring trough visits of growing-finishing pigs with UHF-RFID. *Comput. Electron. Agric.* **2018**, *144*, 144–153. [CrossRef]
2. Wang, R.; Gao, R.; Li, Q.; Dong, J. Pig Face Recognition Based on Metric Learning by Combining a Residual Network and Attention Mechanism. *Agriculture* **2023**, *13*, 144. [CrossRef]
3. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef] [PubMed]
4. Weinstein, B.G. A computer vision for animal ecology. *J. Anim. Ecol.* **2018**, *87*, 533–545. [CrossRef]
5. Li, X.; Lei, Y.K. Radiation source individual identification using machine learning method. In Proceedings of the 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing China, 24–26 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1001–1005.
6. Wang, W.; Wu, J.; Yu, H.; Zhang, H.; Zhou, Y.; Zhang, Y. A Review of Animal Individual Recognition Based on Computer Vision. In Proceedings of the International Conference of Pioneering Computer Scientists, Engineers and Educators, Chengdu, China, 19–22 August 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 287–309.
7. Hou, J.; He, Y.; Yang, H.; Connor, T.; Gao, J.; Wang, Y.; Zeng, Y.; Zhang, J.; Huang, J.; Zheng, B.; et al. Identification of animal individuals using deep learning: A case study of giant panda. *Biol. Conserv.* **2020**, *242*, 108414. [CrossRef]
8. Marsot, M.; Mei, J.; Shan, X.; Ye, L.; Feng, P.; Yan, X.; Li, C.; Zhao, Y. An adaptive pig face recognition approach using Convolutional Neural Networks. *Comput. Electron. Agric.* **2020**, *173*, 105386. [CrossRef]
9. Chen, C.; Zhu, W.; Norton, T. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Comput. Electron. Agric.* **2021**, *187*, 106255. [CrossRef]
10. Brünger, J.; Gentz, M.; Traulsen, I.; Koch, R. Panoptic segmentation of individual pigs for posture recognition. *Sensors* **2020**, *20*, 3710. [CrossRef] [PubMed]
11. Zhang, L.; Gray, H.; Ye, X.; Collins, L.; Allinson, N. Automatic individual pig detection and tracking in pig farms. *Sensors* **2019**, *19*, 1188. [CrossRef] [PubMed]
12. Xiao, D.; Feng, A.; Liu, J. Detection and tracking of pigs in natural environments based on video analysis. *Int. J. Agric. Biol. Eng.* **2019**, *12*, 116–126. [CrossRef]
13. Gan, H.; Guo, J.; Liu, K.; Deng, X.; Zhou, H.; Luo, D.; Chen, S.; Norton, T.; Xue, Y. Counting piglet suckling events using deep learning-based action density estimation. *Comput. Electron. Agric.* **2023**, *210*, 107877. [CrossRef]
14. Liu, C.; Su, J.; Wang, L.; Lu, S.; Li, L. LA-DeepLab V3+: A Novel Counting network for pigs. *Agriculture* **2022**, *12*, 284. [CrossRef]
15. Jianquan, L.; Xiao, W.; Yuanlin, N.; Ying, Y.; Gang, L.; Yang, M. Detection of Herd Pigs Based on Improved YOLOv5s Model. *Int. J. Adv. Comput. Sci. Appl.* **2023**, *14*. [CrossRef]
16. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
17. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
18. He, Y.; Zhu, Y.; Li, H. Cross-layer channel attention mechanism for convolutional neural networks. In Proceedings of the Thirteenth International Conference on Digital Image Processing (ICDIP 2021), Singapore, 20–23 May 2021; SPIE: (Singapore) 2021; Volume 11878, pp. 437–444.

19. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1483–1498. [CrossRef] [PubMed]
20. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
21. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Proceedings, Part V 13; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
22. Wu, Y.; Wu, B.; Zhang, Y.; Wan, S. A novel method of data and feature enhancement for few-shot image classification. *Soft Comput.* **2023**, *27*, 5109–5117. [CrossRef]
23. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
24. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
25. Sharma, S.; Sharma, S.; Athaiya, A. Activation functions in neural networks. *Towards Data Sci.* **2017**, *6*, 310–316. [CrossRef]
26. Fushiki, T. Estimation of prediction error by using K-fold cross-validation. *Stat. Comput.* **2011**, *21*, 137–146. [CrossRef]
27. Zhai, X.; Tian, J.; Li, J. Instance segmentation method of adherent targets in pig images based on improved mask R-CNN. In Proceedings of the 2021 33rd Chinese Control and Decision Conference (CCDC), Kunming, China, 22–24 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 368–373.
28. Wang, F.; Fu, X.; Duan, W.; Wang, B.; Li, H. Visual Detection of Lost Ear Tags in Breeding Pigs in a Production Environment Using the Enhanced Cascade Mask R-CNN. *Agriculture* **2023**, *13*, 2011. [CrossRef]
29. Xie, Q.; Wu, M.R.; Bao, J.; Yin, H.; Liu, H.; Li, X.; Zheng, P.; Liu, W.; Chen, G. Individual pig face recognition combined with attention mechanism. *Trans. Chin. Soc. Agric. Eng* **2022**, *38*, 180–188.
30. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.