Han *et al*. *Intell Robot* 2024;4(1):20-38
DOI: 10.20517/ir.2024.02

Intelligence & Robotics

Research Article

Open Access

Pig-ear detection from the thermal infrared image based on improved YOLOv8n

Hui Han[1,2,#], Xianglong Xue[2,#], Qifeng Li[2], Hongfeng Gao[1], Rong Wang[2], Ruixiang Jiang[2], Zhiyu Ren[2], Rui Meng[2], Mingyu Li[2], Yuhang Guo[2], Yu Liu[2], Weihong Ma[2]

[1]College of Information Engineering, Henan University of Science and Technology, Luoyang 471023, Henan, China.
[2]Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, China.
[#]Authors contributed equally.

Correspondence to: Dr. Weihong Ma, Dr. Qifeng Li, Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing Agricultural Science Building, No.11, Shuguang Garden Middle Road, Haidian District, Beijing 100097, China. E-mail: mawh@nercita.org.cn; liqf@nercita.org.cn

How to cite this article: Han H, Xue X, Li Q, Gao H, Wang R, Jiang R, Ren Z, Meng R, Li M, Guo Y, Liu Y, Ma W. Pig-ear detection from the thermal infrared image based on improved YOLOv8n. *Intell Robot* 2024;4(1):20-38. http://dx.doi.org/10.20517/ir.2024.02

Received: 8 Nov 2023 **First Decision:** 26 Dec 2023 **Revised:** 23 Jan 2024 **Accepted:** 25 Jan 2024 **Published:** 31 Jan 2024

Academic Editor: Simon X. Yang **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

In the current pig scale breeding process, considering the low accuracy and speed of the infrared thermal camera automatic measurement concerning the pig body surface temperature, this paper proposes an improved algorithm for target detection of the pig ear thermal infrared image based on the YOLOv8n model. The algorithm firstly replaces the standard convolution in the CSPDarknet-53 and neck network with Deformable Convolution v2, so that the convolution kernel can adjust its shape according to the actual situation, thus enhancing the extraction of input features; secondly, the Multi-Head Self-Attention module is integrated into the backbone network, which extends the sensory horizons of the backbone network; finally, the Focal-Efficient Intersection Over Union loss function was introduced into the loss of bounding box regression, which increases the Intersection Over Union loss and gradient of the target and, in turn, improves the accuracy of the bounding box regression. Apart from that, a pig training set, including 3,000 infrared images from 50 different individual pigs, was constructed, trained, and tested. The performance of the proposed algorithm was evaluated by comparing it with the current mainstream target detection algorithms, such as Faster-RCNN, SSD, and YOLO families. The experimental results showed that the improved model achieves 97.0%, 98.1% and 98.5% in terms of Precision, Recall and mean Average Precision, which are 3.3, 0.7 and 4.7 percentage points higher compared to the baseline model. At the same time, the detection speed can reach 131 frames per second, which meets the requirement of real-time detection. The research results show that the improved pig ear detection method based on YOLOv8n proposed in this paper can accurately locate the pig ear in thermal infrared images and provide a reference and basis for the subsequent pig body temperature detection.

Keywords: Thermal infrared image, YOLOv8n, target detection, convolution, MHSA, loss function

1. INTRODUCTION

1.1. Backgrounds

The pig farming industry occupies an important position in livestock and poultry farming, and its development is crucial for the prosperity of the economy and people's livelihood. During the process of pig farming, the existence of various diseases, including swine fever, swine pneumonic disease, swine influenza, transmissible chest, blue ear disease, and other diseases, is a major challenge for the pig industry[1], and they are usually reflected in the ear temperature of the pig to varying degrees, which can be detected by testing the temperature of the pig ear to understand the changes in the state of pig health[2,3]. At present, small and medium-sized pig barns in China still mainly rely on manual observation of disease inspection, which consumes a lot of human resources, cannot realize the timely, comprehensive detection of all pigs, and suffers a risk of cross-infection between humans and animals. Therefore, efficient, portable, real-time temperature detection of pigs has become a hotspot of concern and research in the field of pig breeding[4,5]. Along with the continuous development and increasing maturity of automated temperature measurement technology, contact surface temperature measurement with thermocouples[6,7] and thermistors, implantable temperature measurement with digital temperature sensors or microchips as the main components[8,9], and non-contact surface temperature measurement with a thermal infrared imager and laser have gradually become the means of livestock and poultry temperature monitoring[10–12]. Both contact and implantable temperature measurements are invasive methods, which, to some extent, cause stress and tension in individual livestock and poultry. Thermal infrared, ultrasonic, laser, and other non-contact temperature measurement methods are the most widely used in the field of temperature measurement due to the advantages of rapid detection and no stress. Among them, as ultrasonic and laser ones are mostly used in aerospace, metallurgy, and other fields, they are not very suitable for the detection concerning the body temperature of livestock and poultry. Thermal infrared temperature measurement in the early days relies on the infrared temperature sensors of livestock and poultry. However, with the continuous advancement of technology, the infrared thermal image can be adopted to detect the infrared thermal radiation emitted by objects when photoelectric conversion and signal processing will be mapped on the surface of the object heat distribution data imaging. In terms of individual animal temperature measurement, the infrared thermal image generated by the infrared thermal imaging camera can be the more intuitive display of the temperature distribution of livestock and poultry and contour information and can be performed in the major intensive farming-type farms in the body temperature patrol collection[13]. Therefore, for the field of livestock and poultry farming, thermal infrared technology, as the core of the non-contact temperature measurement, has a broader application prospect. However, to automatically detect the body temperature of pigs through thermal infrared technology, the most critical part is to realize the accurate identification of the temperature measurement parts on the body surface of pigs, and thus, it is of great significance to develop a high-precision automatic detection of the pig ear.

1.2. Challenges

Recently, computer vision technology and thermal infrared technology have been gradually applied to various scenarios in the field of modern livestock and poultry breeding by virtue of their innate advantages of non-contact, low cost, and high efficiency. Among them, there is no lack of relevant explorations on automatic detection of pig temperature. For example, Lu *et al.* used a support vector machine to identify pig heads in infrared images, achieving a certain degree of automatic identification[14]. According to Zhang, the ear root region of each pig in the image was automatically cropped out by calibrating the grey value range and image dissimilarity operation, thus obtaining the grey value of the ear root, and then calculating the average temperature of the ear root through the T-G algorithm to achieve the automatic detection of pig body temperature[15]. Although the above method can complete the automatic identification and extraction of the pig temperature,

there are problems. For example, the detection method is too complicated, incomplete in one step, and easily affected by the pig breeding environment regarding the part identification effect[16,17].

Fortunately, in recent years, deep learning, as one of the most effective image detection methods, has been widely used in livestock and poultry related research fields[18]. These applications include individual animal detection[19,20], behavior discrimination[21], body condition analysis[22], and disease diagnosis[23]. With the continuous development of deep learning, target detection methods derived from convolution networks can achieve recognition and localization of target areas, which can better satisfy the needs of livestock and poultry individuals in part detection[24,25]. Particularly, a deep learning model developed based on infrared thermal imaging can be adopted for the automatic detection of mastitis in dairy cows, with a detection accuracy of 83.33%[26]. A teat detection model based on R2Faster R-CNN enables automatic milking of cows with an average detection accuracy of 80.41% for the model[27]. The YOLOv3 network based on deep learning can also achieve high-precision detection of key parts such as the legs, back, and head of cows in natural scenes[28]. Therefore, the target detection method is also gradually applied in the research on the automatic detection of pig temperature by the infrared thermal image. Liu used a two-stage target detection model to explore the temperature inspection of pigs in a facility farm using infrared thermal imaging[29]. Faster-RCNN was employed to detect the temperature in the ear region of pigs with an average detection accuracy of 85.4%. Ma *et al.* proposed an improved geometric contour model-based algorithm for segmenting adhered pigs in group pig farming[30]. The algorithm achieved an accuracy rate of 98% in correctly segmenting the adhered areas, providing technical support for intelligent pig farming. Zhu *et al.* introduced a technique for detecting the pig ear region using a combination of infrared and visible light images[31]. Following the method, an enhanced active shape model is utilized to automatically detect the segmented ear region. By comparing the automatically detected ear region with the manually segmented ear region, it was shown that 84% of the detected regions had an overlap percentage greater than 0.8, indicating a favorable detection performance. Zhou *et al.* put forward an improved maximum interclass variance algorithm based on thermal infrared images[32]. The method can detect 100% of the complete pig images with ear root features, but it cannot deal with the situation that some ear features are obscured. Huang *et al.* used a deep learning algorithm model based on YOLOv5 and multi-target tracking for automatic identification and counting of pigs, which only has high accuracy for identifying and counting pigs under ten heads and cannot be well applied to the large-scale pig sheds[33].

Although the accuracy of these pig temperature detection methods based on target detection models is high, the models, such as Faster-RCNN and Yolov5, have problems including large model size, a large number of parameters, and low real-time performance, which leads to the high requirements of the detection model for the hardware platform, and it is difficult to be transplanted to mobile or embedded terminal devices for practical applications. As a classic one-stage target detection model, Yolov8n is characterized by the smallest model size and the smallest number of parameters among the five versions of Yolov8. In this case, it is easier to be embedded in edge devices[34]. In addition, the convolutional layer module can extract feature information in the image, hence effectively improving the accuracy of model detection. Due to the attention mechanism module, the model can focus on the expression of valid feature information and attenuate the interference of invalid feature information. As a key step in target localization performance, the bounding box regression loss function can accelerate the convergence speed of the model, thus enhancing the target localization accuracy.

1.3. Objectives

With the proposal of the deformable convolutional layer module and bounding box regression loss function, as well as the attention to vital features by the attention mechanism module, it is necessary for this study to improve the lighter weight pig temperature site detection model Yolov8n with the help of these innovative techniques, and thus, it can achieve stronger robustness, higher accuracy, and better real-time performance, which can be applied to satisfy the portable pig temperature inspection in the intensive farming scenario.

2. IMPROVEMENT ALGORITHM FOR TARGET DETECTION BASED ON YOLOV8N

The YOLOv8n target detection model consists of three main components: the backbone network, the neck network, and the predictive output header. The backbone network of the YOLOv8n model employs an enhanced CSPDarkNet53 structure, which down-samples the input features to generate five different scale features. Figure 1A illustrates the structure of the backbone network, which employs the Cross-Scale Part (CSP) module in the original design. However, in this study, the Deformable Cross-Stage Partial Network fusion (DC2f) module is involved instead. The DC2f module is connected by a gradient shunt to improve the information flow of the feature extraction network while maintaining computational efficiency [Figure 1B]. The Deformable Cross-Stage Partial Network (DCBS) module uses a convolution operation on the input data [Figure 1C], followed by batch normalization and activation using the Sigmoid Linear Unit (SiLU). To achieve adaptive-size output, the backbone network incorporates the Spatial Pyramid Pooling Fast (SPPF) module, which pools the input feature maps onto a fixed-size map. Different from the Spatial Pyramid Pooling (SPP) structure[35], SPPF reduces computational effort and latency by sequentially connecting the three largest pooling layers [Figure 1D].

Incorporating the Path Aggregation Network (PAN) concept, YOLOv8n uses a PAN-Feature Pyramid Network (PAN-FPN)[36,37] structure in the neck, as presented in Figure 1E. Unlike the neck structures of YOLOv5 and YOLOv7 models, YOLOv8n removes the convolution operation following up-sampling in the PAN structure, causing a lightweight model without sacrificing performance. PAN-FPN employs top-down and bottom-up network structures to leverage the combination of shallow positional and deeper semantic information through feature fusion, which can thus ensure the diversity and comprehensiveness of features.

The head layer of YOLOv8n utilizes a decoupled head structure, as depicted in Figure 1F. This structure is composed of two separate branches for object classification and predictive bounding box regression, each with its own distinct loss functions. The classification task employs Binary Cross-Entropy (BCE) loss, while the bounding box regression task utilizes Distributed Focus Loss (DFL)[38] and Complete Intersection Over Union (IOU) (CIOU)[39]. This detection structure enhances detection accuracy and speeds up model convergence. YOLOv8n is an unanchored detection model that precisely defines positive and negative samples. It also employs a task-aligned allocator for dynamic sample assignment, further enhancing the detection accuracy and robustness of the model.

In this paper, an enhanced target detection algorithm called YOLOv8n-DMF [YOLOv8n-Deformable Convolution v2 (DCv2)+ Multi-Head Self-Attention (MHSA)+ Focal-Efficient Intersection Over Union (Focal-EIOU)] based on the YOLOv8n network model is introduced. The algorithm is specifically designed for fast and accurate detection of pig ears in a pig house environment. To address the limitations of traditional convolution kernels, the algorithm replaces them with DCv2[40], a more adaptive and generalizable alternative, in both the backbone network and the neck network, which is conducive to better adapting to unknown changes for the model. In addition, to improve the expressive power of the model and expand the sensory field of view of the backbone network, the MHSA module is integrated into the backbone network[41]. This integration enhances the model's ability to capture relevant information. The algorithm introduces the Focal-EIOU loss function[42] in the bounding box regression task in order to overcome the limitations of the CIOU-based loss function, which obviously results in slow convergence and inaccurate regression outcomes. All these improvements are visually represented in the red dashed box in Figure 1.

2.1. The deformable convolution

In the domain of visual recognition, one of the primary hurdles is effectively dealing with geometric variations or capturing geometric transformations in the object scale, pose, viewpoint, and part deformation. Conventional modules used for visual recognition face a constraint of fixed geometry, where the convolutional unit samples the input feature map at a fixed position, lacking an internal mechanism to handle geometric trans-

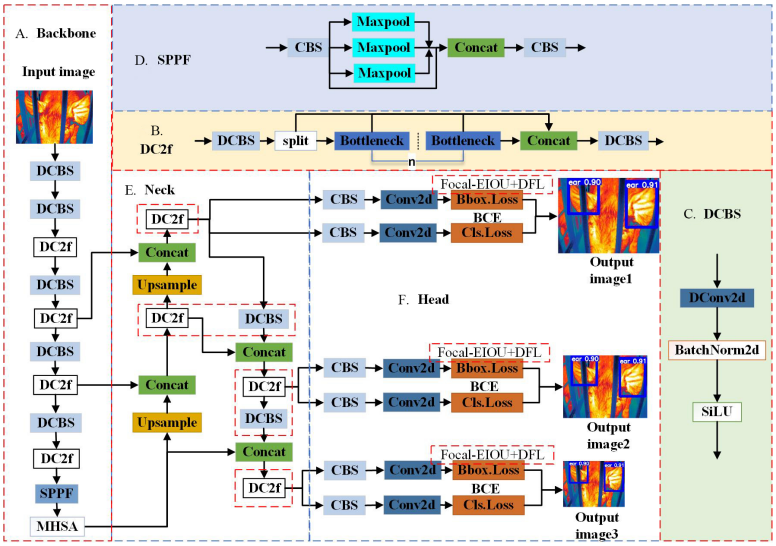

Figure 1. Structure of the YOLOv8n-DMF model.

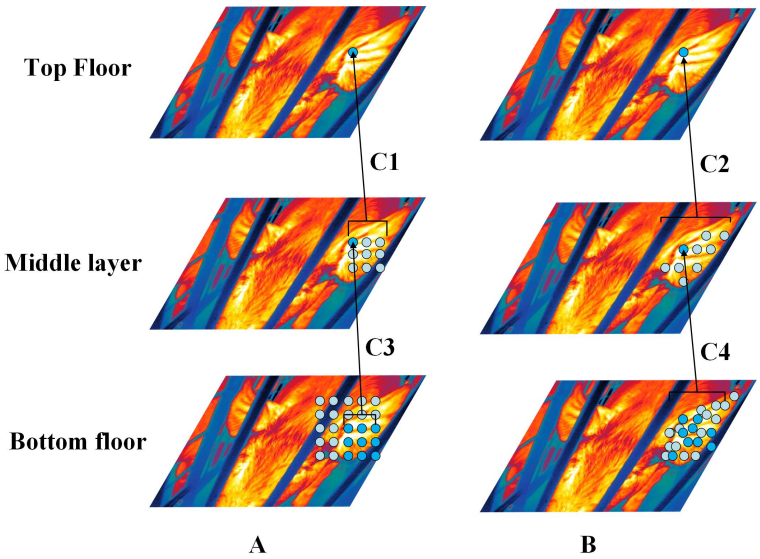

Figure 2. Comparison of feature sampling for (A) standard convolution and (B) deformable convolution.

formations. To overcome the existing limitations, in this study, deformable convolution (DC), which enables the convolution kernel to adaptively adjust its own shape, is put forward. This approach extracts more key information from the feature map, causing improved performance. The differences between DC and standard convolution are shown in Figure 2, providing an intuitive understanding of the proposed method.

Figure 2 illustrates two mappings of 3 × 3 convolutional layers from the bottom to the middle layer and from the middle layer to the top layer, as depicted in (A) standard convolution and (B) DC. The two 3 × 3 convolutional layer mappings shown in (A) are standard 3 × 3 square samples, and (B) shows non-standard shaped samples, but the sampled points are still 3 × 3, in line with the generalized definition of 3 × 3 convolution. The top images in (A) and (B) represent activation units corresponding to distinct objects, whereas the middle layer displays the sampling region of the top activation unit. Similarly, the bottom layer is the sampling region of the activation unit in the middle layer. The (C1-C4) represents the convolution operation, and the convolution

kernel corresponds to nine points and nine counts in the sampling region. Then, sampling is conducted to obtain one point of the activation unit on top of it, i.e., one value.

As shown in Figure 2, it becomes obvious that (A) standard convolution fails to extract complete features of the pig's left ear due to its inability to consider the variations in shape and size across different objects during feature point sampling. In addition, (B) DC considers the deformations of the target, resulting in the majority of the mapped sampling points in the lower layer covering the target. In this case, sampling the complete features of the irregular pig's left ear can be achieved. The operation of standard convolution can be broken down into two distinct steps:

(1) Sample the input feature map x using the 3×3 regular grid R ($R = \{(-1,-1),(-1,0),\ldots,(0,1),(1,1)\}$); (2) Perform a combined sum of sampling values weighted by w.

Usually, a standard convolution defines a 3×3 kernel with an expansion of 1 (the kernel is a 2-dimensional matrix, length × width). The formula given in Equation (1) represents the calculation for each position p_0 on the output feature map y

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{1}$$

where p_n enumerates the position within the output feature map.

DC adds an offset $\{\Delta p_n \mid n = 1,\ldots,N\}$ to the standard convolution operation, where $N = |R|$. By this offset, the standard convolutional sampling becomes a deformable convolutional one. Then, DC can be written as

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{2}$$

where data samples are acquired at non-uniform offset positions, denoted as $p_n + \Delta p_n$. Considering that the offset Δp_n is commonly represented in fractional or decimal form, Equation (2) is executed by employing the method of bilinear interpolation. Then, Equation (3) is given as

$$x(p) = \sum_{q} G(q,p) \cdot x(q) \tag{3}$$

where p indicates a variable (fractional) position ($p = p_0 + p_n + \Delta p_n$), while q denotes all integer positions within the feature map x.

Additionally, the variable $G(\cdot,\cdot)$ denotes a two-dimensional bilinear interpolating kernel. Notably, the kernel G is two-dimensional and can be divided into two separate one-dimensional kernels. The equation is obtained by

$$G(q,p) = g(q_x,p_x) \cdot g(q_y,p_y) \tag{4}$$

where $g(a,b) = \max(0, 1-|a-b|)$. Equation (3) is fast to compute because it is non-zero for only a few q.

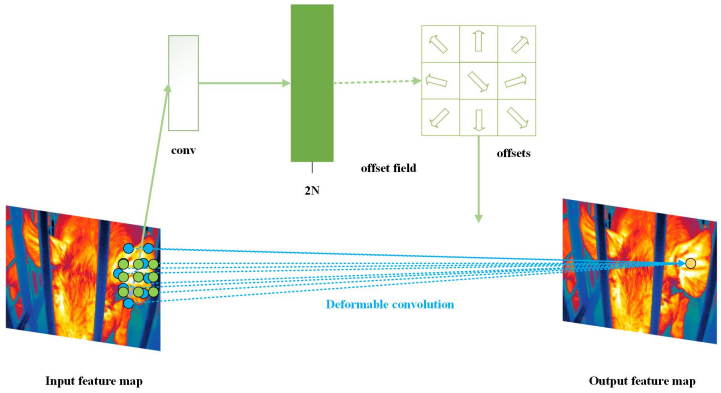

Figure 3. Schematic diagram of 3 × 3 deformable convolution.

As described in Figure 3, the DC process begins similarly to standard convolution. Initially, the feature map is extracted from the input image using a standard convolution kernel. Then, the obtained feature map serves as input for another convolution layer, which generates the offset field responsible for the deformation in DC. This offset layer, with a dimension of $2N$, facilitates translation in the plane by modifying both the x and y values in both directions. During the training phase, the convolution kernel for generating output features and that for producing offsets are simultaneously learned. The offsets are learned through backpropagation, employing an interpolation algorithm.

While DCv1 incorporates DC to enhance adaptability to geometric transformations of the target, it encounters a visualization issue where the corresponding position of the sensory field of view extends beyond the target range. Therefore, the features are still unaffected by the image content, hampering the effectiveness of DCv1. To address the problems of DCv1, this paper adopts DCv2, which extends DC and enhances the modeling capability. Its calculation formula is obtained by

$$y\left(p_0\right)=\sum_{p_n \in R} w\left(p_n\right) \cdot x\left(p_0+p_n+\Delta p_n\right) \cdot \Delta m_n \tag{5}$$

where Δp_n is the learned offset, which allows the sampling points to be irregularly shaped, and Δm_n refers to the learned weight, which solves the problem of the sensory field exceeding the target range in DCv1.

2.2. Efficient self-attention mechanisms

This paper is primarily aimed to investigate the identification of small targets and edge information. This study is conducted within a complex environmental background, which poses significant challenges to the detection model. Specifically, the detection model necessitates a strong capability to identify edge information and effectively suppress background information. To address this concern and enable the detection model to focus on diverse crucial information rather than specific details, the research incorporates the MHSA mechanism into the backbone network of the model. MHSA processes the original input sequences with multiple sets of self-attention, followed by splicing the results of each set of self-attention and performing a linear transformation to get the final output results. This approach not only solves the defect of the self-attention mechanism that overly focuses on its own position but also uses the multi-head self-attention mechanism to give the output of the attention layer containing information about coded representations in different subspaces, thus enhancing the expressive power of the model.

YOLOv8 refers to a Convolutional Neural Network (CNN) model that overcomes a limitation of traditional CNNs. Typically, CNNs perform localized processing, lacking the ability to capture relationships between global features. By contrast, models incorporating attention mechanisms can assess the degree of correlation

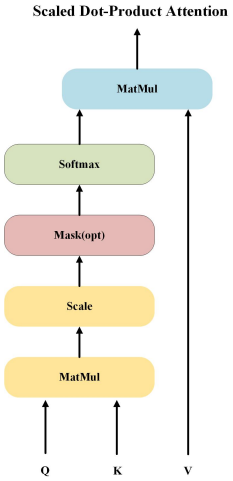

Figure 4. Structure of the scaled dot-product attention mechanism.

between data points within a global perceptual field. By incorporating effective attention mechanisms, robust and powerful data-driven models can be constructed. This increased flexibility enables these models to effectively handle complex and large-scale datasets. The self-attention mechanism operates through a series of steps. Initially, the input data sequence X is encoded into an input matrix $[x_1, x_2, \ldots, x_n]$ that is then linearly transformed using three trainable parameter matrices W^Q, W^K and W^V, resulting in three new matrices, namely Query (Q), Key (K), and key Value (V). The dot product between the queries and corresponding key values is computed, normalized, and multiplied by the matrix V to obtain a weighted sum. The calculation formula for this process is obtained by

$$\text{Scaled Dot-Product Attention } (Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_K}}\right)V \tag{6}$$

where $\sqrt{d_K}$ is used to prevent the gradient of the result from vanishing, and d_K denotes the size concerning the dimension of the matrix K.

From Figure 4, it can be observed that firstly, Q and K^T are subjected to MatMul's dot product operation to generate the similarity matrix. Then, each element in the similarity matrix is divided by $\sqrt{d_K}$, where d_K is the dimension size of K and this division is the Scale step in Figure 4. The Mask step is used to extract the region of interest and mask the interference in the image. Then, the normalization operation of the Softmax step is performed, and each value obtained is a weight coefficient greater than 0 and less than 1 and sums to 0. This result is considered as a weight matrix. Finally, the weighted summation is computed using the obtained weight matrix multiplied by the matrix V.

To address the problem that the YOLOv8n master network exists to encode information about the target location, it will excessively focus on its own location, leading to difficulty in extracting all the key feature information. In this paper, MHSA is proposed and introduced into the backbone network of YOLOv8n. The essence of the mechanism is to process the original input sequence with multiple sets of self-attention, thus obtaining h multi-head outputs. The mechanism can obtain more feature information, which enhances the attention of the model to different features and thus extends the model's network horizon. Its structure is shown in Figure 5.

It has been derived from the principle of the self-attention mechanism. In Figure 4, Q, K, and V are acquired by multiplying the input X with W^Q, W^K, and W^V, respectively; W^Q, W^K, and W^V are trainable parameter

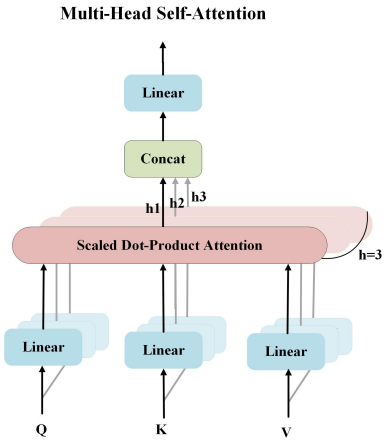

Figure 5. Structure of the multi-head self-attention mechanism.

matrices. As can be seen from Figure 6, there are three output heads ($h1$, $h2$, and $h3$) in the graph. Besides, correspondingly, for the same output X, three different groups of W^Q, W^K, and W^V are defined, and each group is computed to generate different Q, K, and V separately. Subsequently, the normalization process is performed through the Linear layer to accelerate the training process and enhance the generalization ability of the model. Then, the self-attention process is carried out to obtain the three output heads: h1, h2, and h3. Finally, the results of each set of self-attention are spliced together to perform a linear transformation and thus acquire the output result (MHSA). The formula can be described as

$$h_i = \text{ Scaled Dot-Product Attention } \left(QW_i^Q, KW_i^K, VW_i^V\right), \tag{7}$$

$$\text{Multi-Head Self-Attention}(Q, K, V) = \text{Concat}\left(h_1, \ldots, h_i\right) W^O \tag{8}$$

where $h_i \quad (i = \{1, 2, 3\})$ is the result concerning each group of the self-attention process; $W_i^Q \in R^{d_{\text{model}} \times d_q}$, $W_i^K \in R^{d_{\text{model}} \times d_k}$, $W_i^V \in R^{d_{\text{model}} \times d_v}$, $W^O \in R^{d_{\text{model}} \times hd_v}$ is the weight matrix of linear transformation; QW_i^Q, KW_i^K, VW_i^V in (7) are the input matrices regarding each group of self-attention processes, respectively; Concat $(h_1, \ldots., h_h)\, W^O$ in (8) is the step of splicing the outputs of each h_i, then multiplying the spliced outputs by $W^O \in R^{d_{\text{model}} \times hd_v}$, and finally carrying out a step for the further fusion of the results; in addition, each self-attention module is limited to $d_k = d_v = d_{\text{model}} / h = 3$, so $hd_v = d_{\text{model}}$.

2.3. Efficient loss functions for accurate bounding box regression

In target detection, bounding box regression plays a vital role in determining the accuracy of target localization. However, previous loss functions used in bounding box regression often overlook the issue of imbalance, where a large number of anchor boxes featuring minimal overlap with the target box contribute the most to the optimization of bounding box regression. The original model of YOLOv8n introduced the CIOU loss function to address this imbalance problem by considering the overlap area, centroid distance, and aspect ratio, which contributed to significant improvements in both convergence speed and detection accuracy compared to previous loss functions. However, the CIOU loss function does not adequately consider the balance between difficult and easy samples. In addition, as it accounts for aspect ratio, it only reflects the difference in aspect ratio instead of the actual disparity between width and height and their respective confidences. Furthermore, the CIOU loss function does not allow for simultaneous adjustment of anchor box length and width, which hampers convergence. To mitigate these adverse effects, this study proposes the EIOU loss function, which explicitly measures the differences in the three geometric factors of bounding box regression: overlap area,

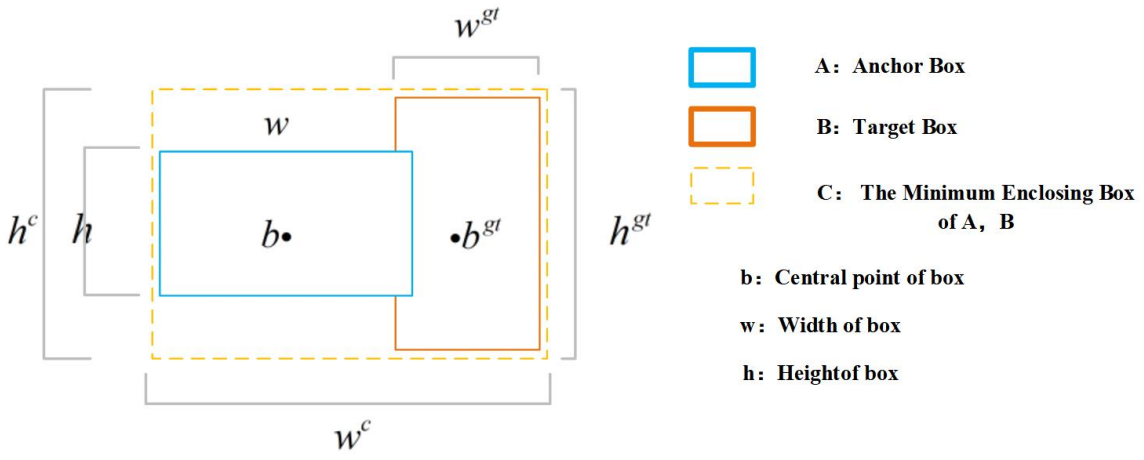

Figure 6. Schematic of the parameters of the efficient loss function for exact bounding box regression.

centroid distance, and aspect ratio. To prioritize high-quality anchor boxes during the regression process, the concept of focal loss is introduced. Finally, the two components are combined to form a new loss function, namely the Focal-EIOU loss function. Its calculation formula is obtained by

$$\begin{aligned} L_{\text{EIOU}} &= L_{\text{IOU}} + L_{\text{dis}} + L_{\text{asp}} \\ &= 1 - \text{IOU} + \frac{\rho^2(b, b^{gt})}{(w^c)^2 + (h^c)^2} + \frac{\rho^2(w, w^{gt})}{(w^c)^2} + \frac{\rho^2(h, h^{gt})}{(h^c)^2} \end{aligned} \tag{9}$$

$$L_{\text{Focal-EIOU}} = \text{IOU}^{\gamma} L_{\text{EIOU}} \tag{10}$$

Where the EIOU loss can be divided into three parts: IOU Loss + Distance Loss + Aspect Ratio Loss. Figure 6 displays the parameters associated with Equation (9). b and b^{gt} are the coordinates concerning the centroid of the target frame and the prediction frame, respectively; $\rho^2(b, b^{gt})$ denotes the Euclidean distance between the two; w^c and h^c are the width and length of the smallest outer rectangle of the prediction frame and the target frame; $\rho^2(w, w^{gt})$ denotes the difference between the width of the target frame and the prediction frame, and $\rho^2(h, h^{gt})$ indicates the difference between the length of the target frame and the prediction frame. To enhance the focus on high-quality samples within the EIOU loss, the concept of Focal loss is incorporated, gaining the final formulation, as given in Equation (10). In this equation, the hyperparameter γ controls the curvature of the curve, and its default value is set to 0.5.

3. EXPERIMENTS

3.1. Experimental data acquisition

This section presents a detailed outline of the experimental methodology utilized in this study, covering various aspects and procedures. Firstly, it outlines the process of data acquisition and dataset creation, detailing the specific methods employed to collect and curate the necessary data for the experiments. Subsequently, the experimental environment, including the hardware and software configurations utilized, and the training strategy implemented to train the models are discussed. Furthermore, the evaluation metrics employed to evaluate and analyze the experimental outcomes are introduced, offering a concise comprehension of the evaluation and comparison process of the models.

The authors are grateful to the Tieqilishi Guanghui Core Pig Farm for the approval of our experiments.

The experimental data for this paper were collected from the Tieqilishi Guanghui Core Pig Farm in Santai

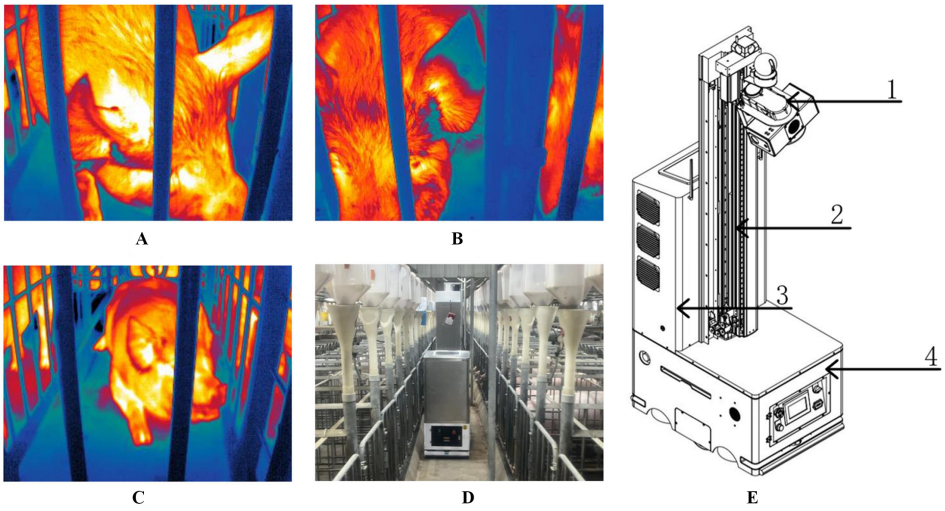

Figure 7. (A)-(C): infrared images at different angles and distances; (D): physical drawing of the inspection robot; (E): design drawing of the inspection robot system. (E)-1: sensing platform; (E)-2: automatic lifting and lowering device; (E)-3: information collection and control system; (E)-4: walking device.

County, Mianyang City, Sichuan Province, China; collections were made from York breeding pigs in restriction pens; the collection period was from April to June 2023, and data collection time was selected from 8:00 am to 10:00 am; the dimensions of the restriction pens in the test barn were 220 cm long, 75 cm wide and 106 cm high. In order to enhance the generalization ability of the model by acquiring diverse data, a selection was made to capture images of live pigs from various angles in an indoor room temperature environment. In terms of images, those captured from close-up overhead angles, close-up side angles, distant overhead angles, and distant frontal angles were involved, resulting in a total of 4,500 thermal infrared images of live pigs with a resolution of 320 × 240 pixels. The acquired infrared images were manually screened at a later stage to remove the images with serious ear loss due to shooting factors, temperature factors and site factors and those without pig ears. In the end, 3,000 thermal infrared images of pig ears were obtained. Some of the infrared images in the dataset and the physical and system diagrams of the acquisition equipment at the test site are shown in Figure 7.

Experimental dataset production

The 3,000 thermal infrared images of pig ears in the dataset were labeled by using horizontal rectangular boxes in the labeling software. The samples were obtained from different individual pigs in 50 pens. This paper tends to identify the ears of live pigs and to avoid the interference of non-target areas to the identification of the target area; the ear is labeled as closely as possible. In the meanwhile, thermal infrared images of pig ears may present challenges, such as obstruction from the pen and variations in ear posture, making it difficult to accurately label the ear in the image. To alleviate this labeling difficulty, pig ear regions in thermal infrared images that are fuzzy, difficult to distinguish, and occluded by over 70% will not be labeled, and the labeling format used is XML. The dataset utilized in this study is partitioned into three subsets, namely training, validation, and test sets, adhering to the widely accepted 8:1:1 ratio employed in mainstream datasets. The training subset comprises 2,400 images, whereas the test and validation subsets consist of 300 images each.

Experimental environment and training strategies

The model training in this study was conducted on a Windows 10 operating system, utilizing an AMD Ryzen 7 5800H processor with Radeon Graphics 3.20 GHz, an NVIDIA RTX 3060 graphics card with 6G of video memory, and 16G of host RAM. The training process employed CUDA version 11.3 and Python version 3.9.12, leveraging the PyTorch deep learning framework with PyTorch version 1.12.1. During the training phase,

optimization was achieved using stochastic gradient descent (SGD) with specific hyperparameters. The initial learning rate was set to 0.01, gradually decreasing to a final learning rate of 0.0001. The cosine annealing hyperparameter was set to 0.1, the momentum factor to 0.937, and the weight decay coefficient to 0.0005. The input size of the images was confirmed as 416 × 416 pixels, with a batch size of 32. During the training process, eight processes were followed, and a total of 300 rounds were conducted.

3.2. Indicators for model evaluation

In this study, the widely accepted evaluation metrics of P, R, and mAP are adopted to assess the training accuracy of the model. These metrics are widely employed in the domain of target detection. The parameter count, computational demands, and weight size of the model were considered to assess its complexity. To evaluate the real-time detection performance of the model, the frame per second (FPS) is utilized as a quantitative measure.

The P metric quantifies the proportion of the predicted algorithm area in relation to the actual detection area. Its algorithm is obtained by

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{11}$$

where T_P (true positive) refers to the number of samples that are accurately predicted as positive by the algorithm. In addition, the F_P (false positive) metric signifies the count of samples that are erroneously classified as positive by the algorithm.

The symbol R represents the ratio of accurately predicted samples to the overall number of positive samples. This metric quantifies the recall or sensitivity of the algorithm in correctly identifying positive instances. Its algorithm is obtained by

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{12}$$

where F_N (false negative) metric denotes the count of samples that were inaccurately classified as negative by the algorithm. This metric indicates the instances where the algorithm failed to identify positive samples correctly.

The term mAP represents the mean average precision (AP), which is calculated as the average value of the individual AP scores. The AP score, in turn, is determined by the area under the precision-recall curve. In the curve, the relationship between the precision and recall values is illustrated, showcasing the model performance in detecting and correctly classifying positive samples. Its algorithm is obtained by

$$mAP = \frac{\sum_{i=1}^{N} \int_0^1 P(R)dR}{N} \times 100\% \tag{13}$$

where N denotes the number of categories. In this study, only one category of pig ear is discussed, so $N = 1$ in this equation.

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Table 1. Results of ablation experiments

Model	Parameters/10^6 M	Computation/GFLOPs	Weight/MB	Precision/%	Recall/%	mAP/%
YOLOv8n	3.20	8.70	6.08	93.7	97.4	93.8
YOLOv8n-D	3.20	8.70	6.15	94.3	97.6	96.2
YOLOv8n-M	3.34	8.90	6.67	96.6	97.0	98.0
YOLOv8n-F	3.20	8.70	6.08	95.7	98.1	95.9
YOLOv8n-DMF	3.34	8.90	6.74	97.0	98.1	98.5

4.1. Ablation experiments

Comparative experiments between the enhanced model and the baseline model, YOLOv8n, are performed to demonstrate the improved detection performance of the enhanced model. Table 1 presents the values of key evaluation metrics for both the improved model and the YOLOv8n model at each stage.

YOLOv8n-D: denotes the replacement of standard convolution with DCv2 in the YOLOv8n backbone and neck networks; YOLOv8n-M: indicates the addition of an MHSA module to the YOLOv8n backbone network; YOLOv8n-F: represents the utilization of a Focal-EIOU loss function to introduce precise bounding box regression in the head network of the YOLOv8n model; YOLOv8n-DMF: signifies the incorporation of DCv2, the MHSA module, and the Focal-EIOU loss function simultaneously into the YOLOv8n network. According to the results from the ablation experiments presented in Table 1, it can be found that each improvement strategy implemented on the YOLOv8n base model contributes to varying extents in enhancing the model's detection performance.

Firstly, due to the introduction of a Focal-EIOU loss function for precise bounding box regression in the original network, model precision, recall, and AP are improved by 2.0, 0.7, and 2.1 percentage points, respectively. Notably, these improvements are achieved without any changes to the number of parameters, computation, and weight size. The Focal-EIOU loss function addresses the issue of gradient vanishing that occurs in the traditional IOU loss when the predicted bounding box does not overlap with the target box. By explicitly quantifying the geometric aspects in the disparities of bounding boxes and integrating the notion of focal loss, the regression procedure prioritizes anchor boxes of superior quality, thereby enhancing accuracy in regression.

Secondly, the inclusion of the MHSA module in the network effectively enhances precision and average accuracy by 2.9 and 4.2 percentage points, respectively, compared to the original YOLOv8n network. However, this enhancement comes at a slight cost of increased parameters, computational effort, and weight size, as well as a slight decrease in recall rate. In the MHSA module, a global multi-head self-attention mechanism is introduced, expanding the receptive field of the main network and improving sensitivity and adaptability to small object detection. The increase in precision is attributed to this expanded sensory field. However, the computational and spatial complexity of the MHSA module leads to a slight increase in model parameters, computation, and weight size.

Lastly, replacing the standard convolution in the original backbone and neck networks with DCv2 yields improvements of 0.6, 0.2, and 2.4 percentage points in model precision, recall, and AP, respectively, compared with the original model. Here, it should be noted that this improvement is achieved without any significant changes in the number of model parameters, computation, and weight size. DCv2 enables the convolution kernel to dynamically adjust its shape based on the input features, resulting in better feature extraction. In addition, the weight term in DCv2 incorporates a penalty mechanism, ensuring that features are not influenced by irrelevant image content and thereby improving network recognition accuracy.

Figure 8 illustrates the variation curves of crucial evaluation metrics for the enhanced models at each stage,

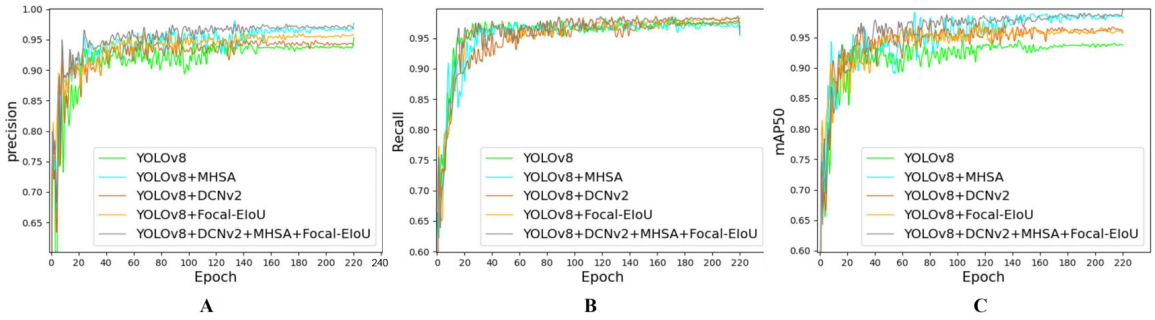

Figure 8. (A) Precision training curves for the improved YOLOv8n model in each stage and YOLOv8n model; (B) Recall training curves for the improved YOLOv8n model in each stage and YOLOv8n model; (C) mAP training curves for the improved YOLOv8n model in each stage and YOLOv8n model

Table 2. Comparison of experimental results of different models

Model	Parameters/10^6 M	Computation/GFLOPs	Weight/MB	mAP/%	FPS
YOLOv5n	7.28	16.70	14.67	90.4	90
YOLOv7n	6.03	13.60	12.15	91.9	105
Faster-RCNN	32.34	378.84	118.06	84.2	17
SSD	21.66	183.76	92.13	78.3	36
YOLOv8n	3.20	8.70	6.08	93.8	148
YOLOv8n-DMF	3.34	8.90	6.74	98.5	131

as well as the YOLOv8n model, throughout the training process. The figure demonstrates that the improved models consistently outperform the YOLOv8n model in terms of precision, recall, and mAP, with the change curves of the YOLOv8n-DMF model exhibiting the most noticeable improvements. All models cease training after approximately 220 training epochs, and the enhanced models at each stage begin to stabilize. Compared with YOLOv8n, the enhanced YOLOv8n-DMF model is featured with faster training speed and superior detection results.

Through comprehensive ablation experiments, it was observed that the improved model resulted in a slight increase in the number of parameters, computation, and weight size. Further analysis revealed that the increased complexity of the network was responsible for enhancing the feature extraction ability of the network, which is an expected outcome. However, in this study, only recognizing small pig ear targets was centered on, and the recognition ability of the model was prioritized. Therefore, the slight increase in the number of parameters, computation volume, and weight size can be disregarded since the improvements in precision, recall, and mAP were significant. With only 3.34 M model parameters, 8.9 GFLOPs of computation, and 6.74 MB of weight size, the improved model achieves 97.0%, 98.1%, and 98.5% in terms of precision, recall, and mAP, respectively, which are 3.3%, 0.7%, and 4.7% higher compared to the original YOLOv8n model, and is able to achieve the goal of enhancing the model recognition accuracy without affecting the deployment of subsequent devices.

4.2. Comparison experiments of different models

Aiming to evaluate the superiority and efficacy of the enhanced algorithm proposed in this research, comparative experiments between multiple widely-used target detection algorithms, including SSD, Faster R-CNN, YOLOv5n and YOLOv7n, and the novel target detection model introduced in this paper, were conducted. The experimental outcomes, as presented in Table 2, offer a comprehensive comparison of the performance achieved by each algorithm.

As can be seen from Table 2, except for the SSD and Faster-RCNN models, which possess lower average ac-

curacy and speed of detection in the ear of live pigs, the other models are characterized by higher average accuracy and faster speed of recognition in the ear of live pigs, which are above 90.0% and 90FPS. Considering that, it is indicated that the Faster-RCNN model and SSD network model are not applicable to the thermal infrared image dataset of live pigs in this paper. Analyzing the reasons, although the SSD model is a one-stage target detection model, which is higher than Faster-RCNN in detection speed, and can also basically satisfy the demand of real-time detection, the model performs detection on low-resolution feature maps, which introduces a relatively large error and leads to poor quality of the detection frame. In terms of the Faster-RCNN, as a two-stage detection model, although the localization of the live pig ear has high accuracy, it is composed of two independent network models, which leads to slower detection speed and is difficult to satisfy the demand of real-time detection. At the same time, it is also shown that the network model of the YOLO series can extract more useful feature information in the ear area of the pig, and the detection model can use these effective features to complete the accurate recognition and localization of the pig ear. For the recognition of the ear root part of the pig, the YOLOv8n-DMF model proposed in this paper achieves 98.5% in the average accuracy of detection, which is 8.1%, 6.6%, 14.3%, 20.2%, and 4.7% higher than that of YOLOv5n, YOLOv7n, Faster-RCNN, SSD, and YOLOv8n models, respectively. In the meanwhile, the YOLOv8n-DMF model also performs well in terms of the number of parameters, computation, model weights, and image detection rate.

Compared with the YOLOv8n model before the improvement, the number of parameters, computational volume, and model weights of the YOLOv8n-DMF model increased slightly, but its average accuracy concerning detecting the temperature-measuring site of the pig ear rose by 4.7%. Furthermore, for real-time detection to be considered effective, it is necessary to achieve a detection rate of more than 24 images per second. As indicated in Table 2, the enhanced network model YOLOv8n-DMF exhibits an impressive detection speed of 131FPS. Notably, this detection speed is only marginally lower than that of the original YOLOv8n model. The reason for the analysis lies in the fact that the replacement of the DC and the addition of the multi-head attention module slightly improve the inference time of the model, but it is still much larger than the minimum requirement of 24FPS and is able to complete the task of real-time detection. In other words, the improved model achieved better detection results without significantly changing the parameters, computational complexity, or weight size of the model and without increasing the demands on high-performance hardware platforms.

In summary, based on the results of the above comparative experiments, improving the detection performance of the model under the premise of ensuring its lightweight makes the YOLOv8n-DMF model proposed in this study easier to deploy on hardware platforms with smaller storage and lower computational performance of edge-based computing, which provides the possibility concerning further exploring the real-time detection system construction at the temperature measurement site on the body surface of a movable pig.

4.3. Visualization analysis

The lack of interpretability in deep learning models poses a significant challenge to their development and practical application. To address this issue, this study aims to provide a more intuitive and convenient way to visualize the detection results of the proposed model. For the purpose of assessing the model performance, its inference effect is analyzed from a detection perspective. Figure 9 illustrates the recognition performance of various target detection models on thermal infrared images of pig ears, which offers a clear comparison of the detection effects achieved by different models.

As can be seen from Figure 9, the improved YOLOv8n-DMF model can accurately identify the ears of pigs with different morphologies in the three pictures under the condition of low-resolution thermal infrared images. Other than that, the highest confidence level of ear identification reaches more than 0.9, while the lowest detection confidence level is also 0.74. In the algorithms of the YOLO series, the confidence level is expressed as the degree of certainty that the model is sure of the detected target. If the value of the confidence level is close to 1, then the model believes that the detection frame contains the target object. Conversely, if the

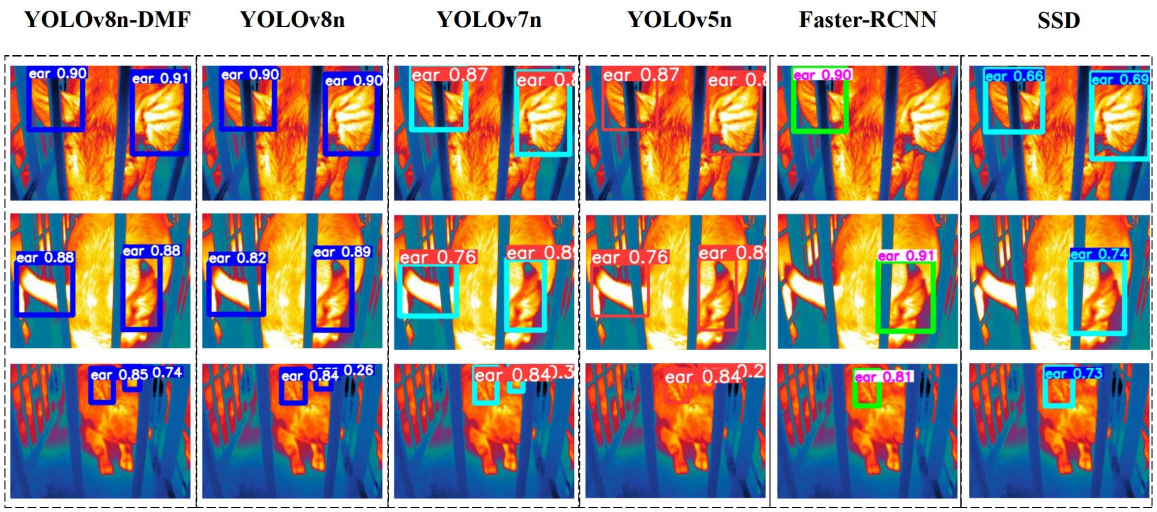

Figure 9. Recognition effect of different models on thermal infrared image of pig ear.

confidence level is close to 0, according to the model, it is assumed that there is no target in the detection box. Therefore, the confidence level can be seen as a probability value indicating the probability of the target existence. Therefore, it can be seen that the improved YOLOv8n-DMF model has high recognition accuracy for pig ears. In addition, the enhanced model is able to detect all the pig ears in the infrared image without any leakage or misdetection. The YOLOv8n, YOLOv7n, and YOLOv5n models are all lower than the improved model in terms of the confidence level for recognizing pig ears in infrared images, although they do not have any leakage or omission in recognizing pig ears in infrared images. The confidence level of the three different models for the occluded pig's ear in the third image is lower than 0.3, which indicates that these models are less capable of recognizing the pig ear in thermal infrared images under complex environments. Although the Faster-RCNN network model is characterized by a target recognition accuracy not inferior to the improved model for all the detected pig ears on the infrared images, its detection results for the three images have the problems of missing detection and omission. In this case, it is shown that the model is less robust and finds it difficult to complete the task of pig ear detection in complex environments. The SSD network model not only has the lowest confidence in recognizing pig ears on infrared pictures but also features the problems of missing detection and omission.

Based on the evaluation of model recognition performance, it is evident that the enhanced YOLOv8n-DMF target detection model, built upon the YOLOv8n framework, exhibits superior performance in accurately recognizing pig ears in infrared images. The model demonstrates exceptional capability in identifying pig ears even when they are partially occluded, without any instances of missed detection or omissions. These results provide concrete evidence proving the effectiveness of the method in practical detection scenarios.

5. CONCLUSIONS

Based on the ablation tests performed on the improved YOLOv8n-DMF model, as well as the results of comparative tests and visualization analyses with other target detection models, the following main conclusions can be drawn.

1. In the ablation experiments, the improved YOLOv8n-DMF target detection model with only 3.34 M parameters, 8.9 GFLOPs of computation, and 6.74 MB of weight size achieves 97.0%, 98.1%, and 98.5% of the model precision, recall, and AP, respectively, which are improved compared to the original YOLOv8n model by 3.3, 0.7, and 4.7 percentage points, indicating that the improved model possesses high recognition accuracy while

maintaining light weight.

2. In the comparison experiments and visualization analysis, the detection effect of the improved YOLOv8n-DMF model is better than that of the YOLOv8n model, and it also performs better than the current mainstream network model. Moreover, the improved target detection model is better for pig ear recognition in infrared images, can accurately identify different shapes and obscured pig ears, and is faster to meet the demand for real-time detection, which verifies the feasibility of the enhanced target detection model for the pig ear in thermal infrared images in this paper, and provides a reference for the next step of pig body temperature detection and mobile deployment.

In this paper, infrared thermal imaging technology and deep learning technology are applied to carry out a series of exploratory studies on non-contact pig body surface thermometry site detection, on the basis of which further in-depth related work can be carried out in the future:

(1) Infrared thermal images have low resolution and lack detailed information about the temperature measurement site compared to visible light images, such as edge details and texture of the ear region. Therefore, it is expected that in future experiments, the detailed features of the temperature measurement site will be increased by fusing the infrared thermal image and the visible light image so as to improve the model performance.

(2) An infrared thermal imaging automatic inspection track can be built in the test barn with a suitable number of pigs and an appropriate environment. Beyond that, the high-precision detection model proposed in this paper should be embedded into the inspection equipment, and an intelligent pig monitoring system with multiple functions, such as temperature warning, pig face recognition, and behavior analysis, should be developed based on the image data automatically collected by the equipment.

DECLARATIONS

Authors' contributions

Made substantial contributions to the conception and design of the study and performed data analysis and interpretation: Han H

Performed data acquisition and provided administrative, technical, and material support: Ma W, Xue X, Li Q, Gao H, Wang R, Jiang R, Ren Z, Meng R, Li M, Guo Y, Liu Y

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the Outstanding Scientist Training Program of Beijing Academy of Agriculture and Forestry Sciences (JKZX202214); Sichuan Science and Technology Program under Grants 2021ZDZX0011; Technological Innovation Capacity Construction of Beijing Academy of Agricultural and Forestry Sciences (KJCX20230204); Beijing Digital Agriculture Innovation Consortium Project (BAIC10-2023).

Conflicts of interest

Ma W is a Junior Editorial Board Member of *Intelligence Robotics*. Other authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

All procedures were performed in accordance with the guidelines recommended by the Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences (Process number AW-2023-04-01). The authors have obtained the approval of the farmer of the Tieqilishi Guanghui Core Pig Farm for the exper-iment.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Albernaz-Gonçalves R, Olmos G, Hötzel MJ. My pigs are ok, why change? – animal welfare accounts of pig farmers. *Animal* 2021;15:100154. DOI
2. Menzel A, Beyerbach M, Siewert C, et al. *Actinobacillus pleuropneumoniae* challenge in swine: diagnostic of lung alterations by infrared thermography. *BMC Vet Res* 2014;10:199. DOI
3. Tzanidakis C, Simitzis P, Arvanitis K, Panagakis P. An overview of the current trends in precision pig farming technologies. *Livest Sci* 2021;249:104530. DOI
4. Zhang Z, Zhang H, Liu T. Study on body temperature detection of pig based on infrared technology: a review. *Artif Intell Agric* 2019;1:14-26. DOI
5. Kammersgaard TS, Malmkvist J, Pedersen LJ. Infrared thermography - a non-invasive tool to evaluate thermal status of neonatal pigs based on surface temperature. *Animal* 2013;7:2026-34. DOI
6. Qu D, Liu S, Wu J, Li Y. Design and implementation of monitoring system for multiple cows body temperature. *Trans Chin Soc Agric Mach* 2016;47:408-12. Available from: https://www.researchgate.net/publication/309529158_Design_and_implementation_of_monitoring_system_for_multiple_cows_body_temperature. [Last accessed on 26 Jan 2024]
7. Li Z. Application of a flexible patch online measurement method in pig body temperature measurement. 2018. (in Chinese). Available from: https://www.zhangqiaokeyan.com/academic-degree-domestic_mphd_thesis/02031264442.html. [Last accessed on 26 Jan 2024]
8. He D, Liu C, Xiong H. Design and experiment of implantable sensor and real-time detection system for temperature monitoring of cow. *Trans Chin Soc Agric Mach* 2018;49:195-202. (in Chinese) DOI
9. Hentzen M, Hovden D, Jansen M, van Essen G. Design and validation of a wireless temperature measurement system for laboratory and farm animals. *Proc Meas Behav* 2012;2012:466-71. Available from: https://archive.measuringbehavior.org/mb2012/files/2012/ProceedingsPDF(website)/Posters/Hentzen_et_al_MB2012.pdf. [Last accessed on 26 Jan 2024]
10. Salles MSV, da Silva SC, Salles FA, et al. Mapping the body surface temperature of cattle by infrared thermography. *J Therm Biol* 2016;62:63-9. DOI
11. Siewert C, Dänicke S, Kersten S, et al. Difference method for analysing infrared images in pigs with elevated body temperatures. *Z Med Phys* 2014;24:6-15. DOI
12. Iyasere OS, Edwards SA, Bateson M, Mitchell M, Guy JH. Validation of an intramuscularly-implanted microchip and a surface infrared thermometer to estimate core body temperature in broiler chickens exposed to heat stress. *Comput Electron Agric* 2017;133:1-8. DOI
13. Giro A, de Campos Bernardi AC, Junior WB, et al. Application of microchip and infrared thermography for monitoring body temperature of beef cattle kept on pasture. *J Therm Biol* 2019;84:121-8. DOI
14. Lu M, He J, Chen C, et al. An automatic ear base temperature extraction method for top view piglet thermal image. *Comput Electron Agric* 2018;155:339-47. DOI
15. Zhang Z. Research on body temperature detection method of breeding pigs based on infrared images. 2021. (in Chinese) DOI
16. Symeonaki E, Arvanitis KG, Piromalis D, Tseles D, Balafoutis AT. Ontology-based IoT middleware approach for smart livestock farming toward agriculture 4.0: a case study for controlling thermal environment in a pig facility. *Agronomy* 2022;12:750. DOI
17. Zheng P, Zhang J, Liu H, Bao J, Xie Q, Teng X. A wireless intelligent thermal control and management system for piglet in large-scale pig farms. *Inf Process Agric* 2021;8:341-9. DOI
18. Bao J, Xie Q. Artificial intelligence in animal farming: a systematic literature review. *J Clean Prod* 2022;331:129956. DOI
19. Zin TT, Pwint MZ, Seint PT, et al. Automatic cow location tracking system using ear tag visual analysis. *Sensors* 2020;20:3564. DOI
20. Zin TT, Misawa S, Pwint MZ, et al. Cow identification system using ear tag recognition. In: 2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech); 2020 Mar 10-12; Kyoto, Japan. IEEE; 2020. pp. 65-6. DOI
21. Lodkaew T, Pasupa K, Loo CK. CowXNet: an automated cow estrus detection system. *Expert Syst Appl* 2023;211:118550. DOI
22. Alvarez JR, Arroqui M, Mangudo P, et al. Body condition estimation on cows from depth images using convolutional neural networks. *Comput Electron Agric* 2018;155:12-22. DOI
23. Zhuang X, Zhang T. Detection of sick broilers by digital image processing and deep learning. *Biosyst Eng* 2019;179:106-16. DOI
24. Wang Y, Kang X, Chu M, Liu G. Deep learning-based automatic dairy cow ocular surface temperature detection from thermal images. *Comput Electron Agric* 2022;202:107429. DOI
25. Wang R, Bai Q, Gao R, et al. Oestrus detection in dairy cows by using atrous spatial pyramid and attention mechanism. *Biosyst Eng* 2022;223:259-76. DOI
26. Zhang X, Kang X, Feng N, Liu G. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. *Comput Electron Agric* 2020;178:105754. DOI
27. Lu Z, Zhao M, Luo J, Wang G, Wang D. Automatic teat detection for rotary milking system based on deep learning algorithms. *Comput Electron Agric* 2021;189:106391. DOI
28. Jiang B, Wu Q, Yin X, Wu D, Song H, He D. FLYOLOv3 deep learning for key parts of dairy cow body detection. *Comput Electron*

Agric 2019;166:104982. DOI
29. Liu Q. Research on pig temperature inspection technology based on thermal infrared image. 2022. (in Chinese) DOI
30. Ma L, Duan Y, Zong Z, Liu G. Segmentation of thermal infrared image for sow based on improved convex active contours. *Trans Chin Soc Agric Mach* 2015;46:180-6. (in Chinese) DOI
31. Zhu W, Liu B, Yang J, Ma C. Pig ear area detection based on adapted active shape model. *Trans Chin Soc Agric Mach* 2015;46:288-95. (in Chinese) DOI
32. Zhou L, Chen Z, Chen D, Yuan Y, Li Y, Zheng J. Pig ear root detection based on adapted otsu. *Trans Chin Soc Agric Mach* 2016;47:228-32,14. (in Chinese) DOI
33. Huang Y, Xiao D, Liu J, Tan Z, Liu K, Chen M. An improved pig counting algorithm based on YOLOv5 and DeepSORT model. *Sensors* 2023;23:6309. DOI
34. Terven J, Cordova-Esparza DM. A comprehensive review of YOLO: from YOLOv1 to YOLOv8 and beyond. 2023. Available from: https://www.researchgate.net/publication/369760111_A_Comprehensive_Review_of_YOLO_From_YOLOv1_to_YOLOv8_and_Beyond. [Last accessed on 26 Jan 2024]
35. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell* 2015;37:1904-16. DOI
36. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21-26; Honolulu, USA. IEEE; 2017. pp. 936-44. DOI
37. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. pp. 8759-68. Available from: https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html. [Last accessed on 26 Jan 2024]
38. Li X, Wang W, Wu L, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems Advances in Neural Information Processing Systems. 2020. pp. 21002-12. Available from: https://proceedings.neurips.cc/paper/2020/file/f0bda020d2470f2e74990a07a607ebd9-Paper.pdf. [Last accessed on 26 Jan 2024]
39. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2020. pp. 12993-3000. DOI
40. Dai J, Qi H, Xiong Y, et al. Deformable convolutional networks. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22-29; Venice, Italy. IEEE; 2017. pp. 764-73. Available from: https://doi.org/10.1109/ICCV.2017.89. [Last accessed on 26 Jan 2024]
41. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. arXiv. [Preprint]. Aug 2, 2023. Available from: https://arxiv.org/abs/1706.03762.
42. Zhang YF, Ren W, Zhang Z, Jia Z, Wang L, Tan T. Focal and efficient IOU loss for accurate bounding box regression. *Neurocomputing* 2022;506:146-57. DOI