Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

A lightweight model of dual - backbone real-time detection transformer for pig body temperature detection and its onsite validation

Jinghan He [a], Hong Zhou [a], Qiuju Xie [a,*], Wenwu Wang [b,*], Xuefei Liu [c], Wenyang Liu [a], Yuhuan Guo [a], Honggui Liu [d,e,f]

[a] *College of Electrical Engineering and Information, Northeast Agricultural University, Harbin 150030, China*
[b] *Centre for Vision, Speech & Signal Processing, Faculty of Engineering & Physical Science, University of Surrey, Guildford, GU2 7XH, UK*
[c] *College of Engineering, Huazhong Agricultural University, Wuhan 430070, China*
[d] *College of Animal Science, Northeast Agricultural University, Harbin 150030, China*
[e] *Engineering Technology Research Center for Intelligent Pig Breeding and Breeding in Northern Cold Regions, Ministry of Agriculture of the People's Republic of China, Harbin 150030, China*
[f] *Key Laboratory of Pig Breeding Facilities Engineering, Ministry of Agriculture and Rural Affairs, Harbin 150030, China*

ARTICLE INFO

Keywords:
Infrared thermography
Temperature detection
Feature extraction
Image alignment
Semantic segmentation

ABSTRACT

Pig body temperature is a critical indicator for health assessment of pigs, and can be detected automatically with machine vision-based models, due to their accuracy and effectiveness. However, it is hard to apply them onsite in large-scale pig farms. A primary issue lies in the absence of lightweight feature extraction models that can effectively address the problem of multimodal information fusion. To fill this gap, this study proposes FP-DETR (Frequency Parallel backbone DEtection TRansformer) by integrating parallel backbone and frequency-spatial domain multimodal fusion technology method into the DETR (DEtection TRansformer) object detection model to compress the model structure and enhance the model's feature extraction capability. The proposed FP-DETR model achieves precision of 98.9% and recall of 96.88% with only 5.9 × 10^6 parameters and 8.8 h training time. Compared with the YOLOv12 model, FP-DETR improves detection speed by 35 FPS and accuracy by 6.3%. FP-DETR's temperature extraction achieved performance with R^2 of 0.957 and mean absolute error (MAE) of 0.108. In addition, equipment development and model integration have been completed, and an on-site experiment has been conducted, showing that the system is about 42.9% faster than manual. Therefore, the proposed model offers excellent performances of efficiency and accuracy as a promising solution for real-time onsite pig body surface temperature detection.

1. Introduction

With the continuous expansion of large-scale pig farming, increasing attention has been directed toward pig health (Tzanidakis, et al., 2021). Body temperature is one of the most important indicators reflecting the health status of pigs (Lu, et al., 2018). Many infectious diseases, such as respiratory infections (Opriessnig, et al., 2011), African swine fever (Salguero, 2020), and porcine reproductive and respiratory syndrome (Benjamin and Yik, 2019), etc., can cause abnormal fluctuations in body temperature. Therefore, the real-time and accurate measurement of pig body temperature is of great significance for the early diagnosis of diseases and the assessment of overall health (Zhang, et al., 2019).

Currently, rectal temperature is widely used as an indicator of pig health in commercial farming (Ramirez and Karriker, 2019). Traditionally, rectal temperature is measured by inserting a mercury or electronic thermometer into the rectum (Cuthbertson, et al., 2019, Sellier, et al., 2014), which is both labor-intensive, time-consuming, and often causes stress in pigs, making it unsuitable for large-scale farms.

With the advancement of digital sensing technologies, implantable biosensors have been explored for temperature measurement. In dairy cows, sensor readings showed a strong correlation with vaginal temperature (r = 0.85) under heat stress conditions (Chung, et al., 2020). In pigs, subcutaneous temperature was about 1°C lower than rectal values but still demonstrated a significant linear relationship (r = 0.88, $P < 0.0001$) (Lohse, et al., 2010). These findings confirm that implantable devices can effectively measure body temperature in animals. However, their high cost and the stress caused by implantation limit their

[*] Correspondence authors.
E-mail addresses: xqj197610@163.com (Q. Xie), w.wang@surry.ac.uk (W. Wang).

https://doi.org/10.1016/j.compag.2025.111234
Received 2 July 2025; Received in revised form 17 September 2025; Accepted 11 November 2025
Available online 1 December 2025

Nomenclature

ASPP	Atrous Spatial Pyramid Pooling
Conv	Convolution
ER	Ear Root
FH	Forehead
$f_1^b(x,y)$	Base layer of the infrared image
$f_1^d(x,y)$	Detail layer of the infrared image
$f_2^b(x,y)$	Base layer of the visible light image
$f_2^d(x,y)$	Detail layer of the visible light image
$f^d(x,y)$	Layers of detail after initial fusion
FPS	Frame per second, frame/s
MAE	Mean absolute error
mAP	Mean Average Precision, %
P	Precision, %
R	Recall, %
RC	Rectum
R^2	Determination coefficient
RMSE	Root mean square error
RoI	Region of interest

practicality in large-scale commercial farming.

In the field of smart agriculture, prior studies have approached digital farming from diverse perspectives. Pratama, et al. (2023) advanced livestock management through a virtual fencing system built on wireless sensor networks and the Haversine method, highlighting spatial monitoring of animal herds, whereas the present study focuses on physiological monitoring via body temperature detection. Hossain and Chowdhury (2024) introduced AgroSense, an internet of things (IoT)-based platform designed to improve crop selection and decision-making, representing a direction toward digitalized agronomy, while this work emphasizes animal health indicators in farming practice. Jumi (2024) concentrated on goat farming by designing an IoT-enabled breeding house, with particular attention to environmental variables such as humidity and gases, in contrast to the current emphasis on body surface temperature as a direct signal of health status. Shofura, et al. (2021) applied artificial neural networks to the classification of monthly weather conditions, showing how artificial intelligence (AI) can improve meteorological forecasting, while here advanced detection transformers are applied to livestock disease early warning. Galina, et al. (2022) integrated Sonic Bloom acoustic stimulation with IoT technology to enhance crop growth, focusing on plant productivity, whereas this study addresses animal health monitoring in real farm environments. Listianingsih and Susanto (2023) proposed frameworks for smart environments and forest cities, emphasizing ecological sustainability at the urban scale, which stands apart from livestock-focused health monitoring at the farm level. Taken together, these works illustrate the diversity of approaches within smart agriculture, while the present research distinguishes itself by combining infrared thermography with a frequency–spatial fusion transformer (FP-DETR) for robust, real-time detection of pig body temperature.

In recent years, infrared thermography (IRT) has gained increasing attention as a non-contact method for measuring animal surface temperature. With its advantages such as convenience, speed, absence of stress responses, and the ability to automate body temperature inspection (Bagavathiappan, et al., 2013), the IRT-based method has been paid more and more attention (Zhang, et al., 2019), especially in inflammation detection (Whittaker, et al., 2023), ovulation monitoring (Marquez, et al., 2019), abnormal behavior recognition and growth assessment (Sasaki, et al., 2016).

In particular, infrared thermography combined with deep learning algorithms has led to significant advances in livestock health monitoring. For example, R^2Faster R-CNN (Lu, et al., 2021) and IT-PETE (Xiao, et al., 2021) have been applied to tasks such as automatic recognition of dairy cow mastitis (Zhang, et al., 2020, nipple detection, and pig's ear detection (Zhou et al., 2017) and temperature extraction. These studies achieved good detection accuracies over 80.41%, time efficiency of 0.19 s and body temperature extraction error of 2.29℃.

Although existing models perform well, they are constrained to spatial domain analysis and struggle to suppress environmental noise such as low-frequency heat sources and high-frequency equipment interference. Relying on pixel-level information makes these models highly sensitive to lighting and background variations, reducing their effectiveness in complex farm environments. In addition, reproducibility across different experimental setups is often limited, which further restricts their reliability and hinders stable deployment in large-scale, noisy farming scenarios.

To overcome these limitations, frequency-domain information and a parallel backbone structure are introduced into the DETR (DEtection TRansformer), a target detection method that directly captures image features through the self-attention mechanism of transformer, to enhance its performance. Specifically, a parallel architecture of Vision Mamba and CNN is adopted to achieve cross-regional correlation of global features through global self-attention modeling and local texture detail extraction, while frequency-domain information is incorporated to decompose high and low-frequency components and construct frequency features. By integrating the Spatial Feature Adaptation (SFA) module and Band Feature Modulation (BFM) module, the proposed FP-DETR (Frequency Parallel backbone DEtection TRansformer) method realizes accurate separation of effective signals from environmental noise. In recently, Frequency Dynamic Convolution Chen, et al. (2025) further demonstrates the value of frequency-domain analysis, highlighting the necessity of multimodal fusion for advancing feature learning.

In a previous study by our research group, the YOLOv5s-BiFPN model was established, developed using infrared thermal imaging to estimate the temperature of six body surface regions in pigs (forehead, eyes, nose, ear roots, back, and anus) (Xie, et al., 2023). Strong correlations were observed between rectal temperature and the temperatures at the ear roots and forehead. Therefore, the ear roots and forehead were selected as the regions of interest (RoIs) for body surface temperature detection.

On this basis, the present study introduces frequency-domain information to enhance image features, separate environmental noise from physiological signals, and integrate a lightweight parallel backbone with additional methods. This design significantly reduces computational complexity while improving the accuracy of feature extraction, thereby enabling robust and efficient pig body temperature detection in real farm environments.

The main contributions of this work are as follows:

(1). A parallel dual-backbone architecture was designed to achieve efficient lightweight feature extraction.
(2). A frequency–spatial fusion strategy was introduced, enhancing feature representation and suppressing environmental noise.
(3). Improved image segmentation and fusion algorithms were developed, enabling high-precision contour recognition while reducing computational cost.
(4). The FP-DETR model was successfully deployed on an inspection robot and validated under real farm conditions, demonstrating its practical feasibility for large-scale pig farming.

2. Materials and methods

2.1. Description of the pig house

The experimental data were collected in winter (February 22 to March 26, 2024) and summer (June 22 to July 31, 2023; July 28 to August 9, 2024) from two locations: Jingzhe Pig Farm in the Yabuli

Forestry Bureau, China (Fig. 1(a)), and HongzhuKangyuan Pig Farm in Harbin City, China (Fig. 1(b)).

At Jingzhe Pig Farm, the pig house was designed with a sloped roof, insulated walls, and adjustable sunshade windows at the top, along with a 1.5 m diameter exhaust fan installed on the outer wall near the entrance. Feeding troughs (2 m × 0.8 m) were placed every five meters along the walkways, and the floor was covered with thick bedding composed of rice husks, rice bran, and corn stalks.

In contrast, the pig house at HongzhuKangyuan Pig Farm had a semicircular vaulted structure with a concrete slatted floor, and its roof was covered with a blue waterproof film to help regulate temperature and humidity. Two exhaust fans (1 m and 1.5 m in diameter) were mounted on the outer wall, and an electronic feeding station was installed inside, with pigs fed twice daily.

2.2. Data collection

Data were collected from 139 ternary hybrid pigs (Duroc × [Landrace × Large White]), aged 240–270 days, with an average weight of 230 kg in summer and 210 kg in winter. Pigs were fed daily at 8:00 and 15:00, and measurements were taken before and after feeding, specifically at 6:00–7:00, 9:00–10:00, and 16:00–17:00. Collected data included skin temperature, rectal temperature, and environmental conditions within the pig houses.

Rectal temperature was measured using a specialized livestock thermometer (Nierni, China; range 20–42.99°C; accuracy ± 0.5°C). The thermometer was inserted 10 cm into the rectum and held for 5–7 s; each measurement was repeated twice, and the mean value was recorded as the rectal temperature. Ambient temperature and humidity were measured using a handheld meter (TA622A, TASI). Body surface temperature was measured with a thermal imaging camera (Fotric Model 287-L20, Fotric, Texas, USA; range 40–150°C; accuracy ± 2°C), with emissivity set to 0.98 and the distance fixed at 1 m, focusing on the ear root and forehead regions.

According to our previous study, the maximum temperature from the ear root and forehead has the highest Pearson's correlation coefficient with porcine rectal temperature (ER: 0.6859, FH: 0.6609), and the maximum temperature is effective in preventing ambient low-temperature interference. So they are selected as the RoIs for pig's body surface temperature detection (Xie, et al., 2023). The collected temperatures on RoIs are shown in Table 1.

2.3. Dataset division

The dataset includes a total of 1688 sets of data, each of which consists of visible light images and corresponding thermal infrared images, with a total of 3376 images. Each image is labeled with Labelme for the ear root and the forehead area. These images are divided into training set (2704 images), validation set (338 images) and test set (338 images) at a ratio of 8:1:1, and the thermal infrared images correspond to the visible light images. In order to ensure the fairness and generalization of the model training, the dataset was randomly divided according to this ratio rather than being split sequentially. To guarantee transparency and reproducibility of the randomization, a fixed random seed was used during the partitioning process.

3. FP-DETR model development

3.1. Workflow of FP-DETR

The lightweight network structure is integrated with the temperature extraction process to improve both the efficiency and accuracy of detection. The overall workflow of FP-DETR is illustrated in Fig. 2.

(1) Image input: Infrared thermal images and corresponding visible light images are used as inputs.

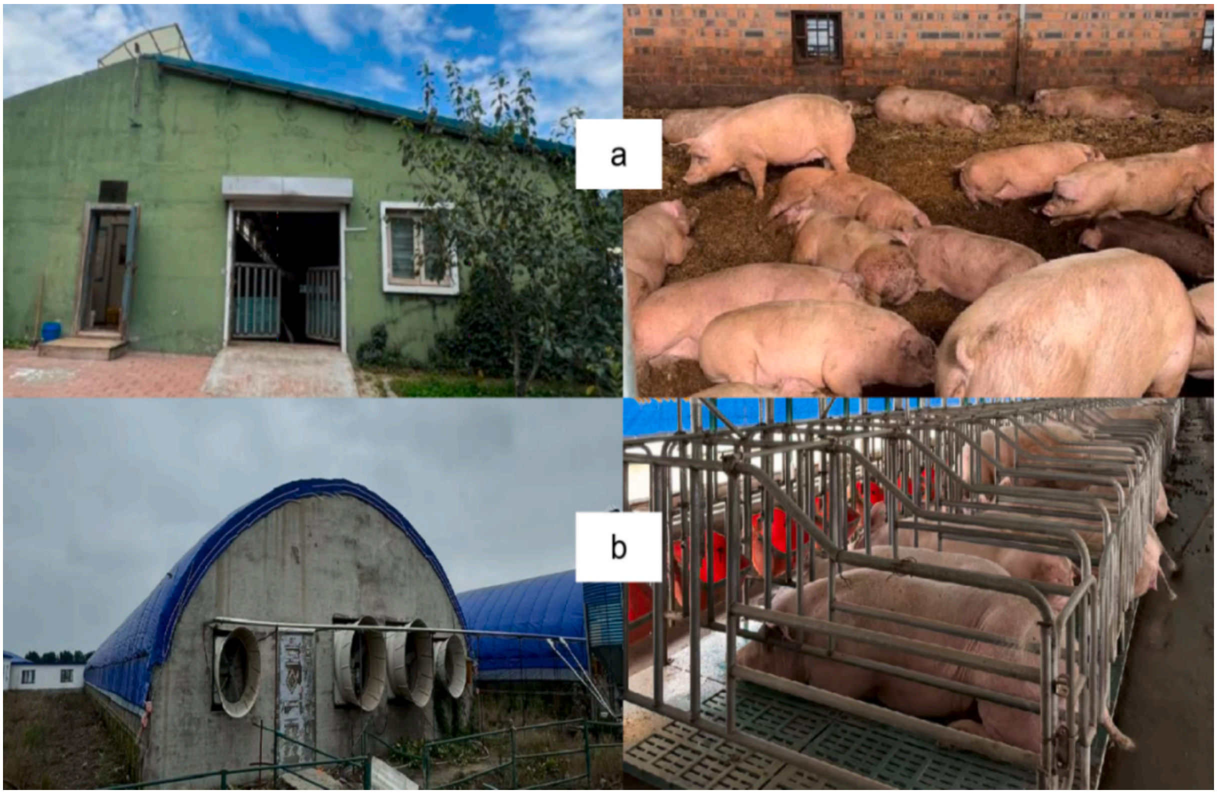

Fig. 1. Pictures of pig houses inside and outside: a. Jingzhe pig farm. b. HongzhuKangyuan pig farm.

Table 1
Experimental temperature statistics.

Area	Summer			Winter		
	Maximum (°C)	Minimum (°C)	Average (°C)	Maximum (°C)	Minimum (°C)	Average (°C)
ER	40.4	29.4	36.8	39.8	25.7	36.6
FH	39.8	25.7	37.7	39.5	29.4	37.2
RC	39.7	37.5	38.5	39.7	37.5	38.5

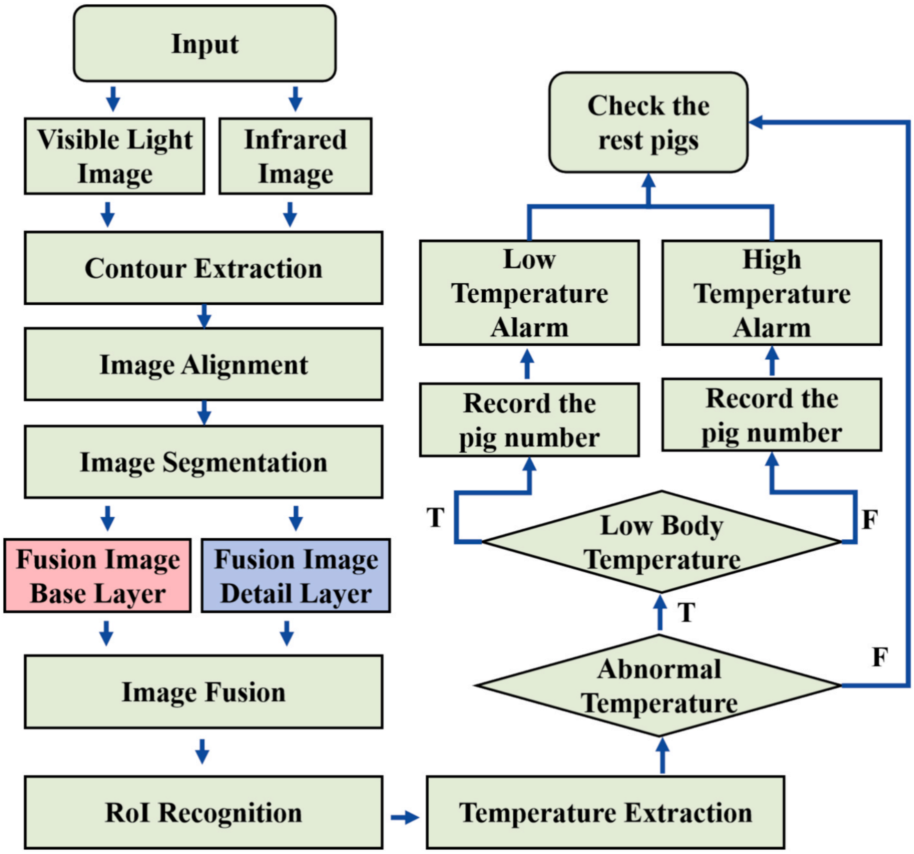

Fig. 2. The overall workflow of FP-DETR.

(2) Image segmentation: Key features on the pig's body surface are extracted from the infrared images, which are then used to segment the image.
(3) Image Fusion: The pig's outline is fused using a trained model. Temperature data from the infrared image are combined with details from the visible light image to generate a fused image containing both temperature and coordinate information.
(4) Temperature extraction: The fused image is passed into the RoI detection module to identify the regions of interest. Based on these regions, the maximum temperature values are extracted from the fused image and recorded.

3.2. *Image segmentation*

3.2.1. *Image registration*

The visible light image captured by the thermal imaging camera has a resolution of 600 × 1200, while the infrared image has a resolution of 512 × 384, resulting in differences in both resolution and visual appearance (Fig. 3). Such discrepancies may affect the accuracy of object recognition.

Therefore, in this study, the AKAZE method was applied to align the contours of visible and infrared images at the key regions of the pig's body. The main steps are as follows:

(1) Grayscale conversion: Both infrared and visible light images were converted to grayscale to reduce computational complexity.
(2) Scale space construction: A scale space was generated through nonlinear diffusion filtering, and key points were detected using a regional method. These key points correspond to local extrema in the image, representing pixels with maximum or minimum gray values within their neighborhoods.
(3) Binary descriptor generation: For each pixel, a binary descriptor was created based on the gradient directions of its neighboring pixels relative to a threshold direction. The value was set to 1 if the gradient direction matched the threshold direction and 0 otherwise.
(4) M-LDB feature description: The Modified Local Difference Binary (M-LDB) operator was used to describe the area surrounding each feature point, as shown in Eq. (1). This operator generates a binary code by comparing the intensity of neighboring pixels with that of the central pixel, thereby capturing local texture information effectively.

$$\mathrm{M}_{\mathrm{LDB}(i,j)} = \sum_{k=0}^{N-1} b_k \times 2^k \tag{1}$$

where *(i,j)* represents the pixel position, *N* is the length of the binary

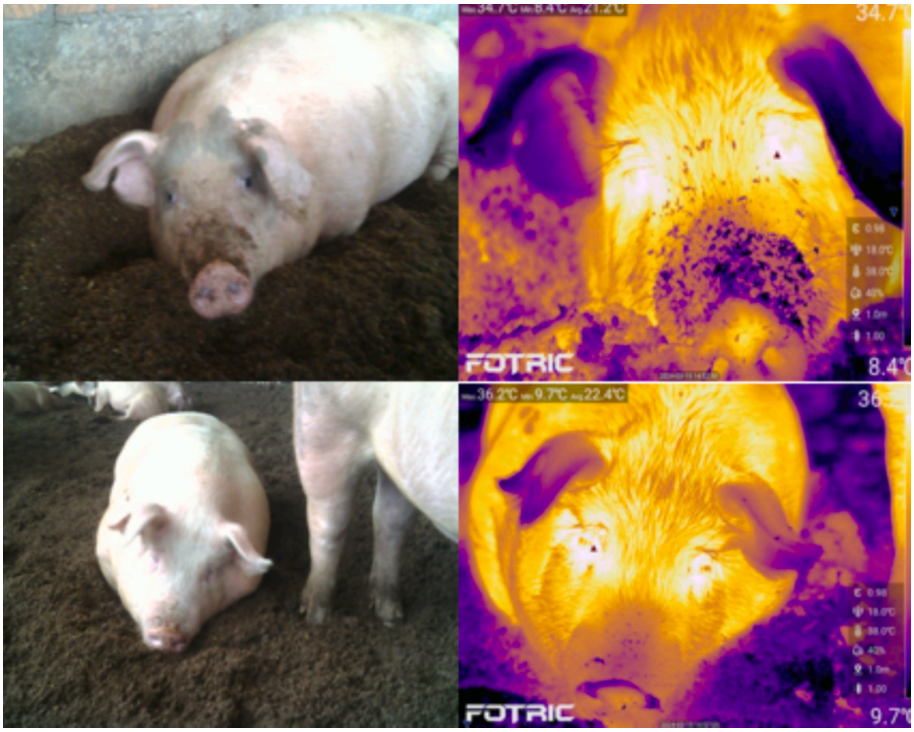

Fig. 3. Visible light image and corresponding infrared image.

descriptor, and b_k is the k-th position in the binary descriptor.

After the feature points were obtained, a transformation matrix was constructed using the Random Sample Consensus (RANSAC) algorithm to scale the visible light image and align it with the target region in the thermal infrared image as shown in Eq. (2). Subsequently, an external rectangular frame was generated based on the body surface contour identified in the thermal infrared image, and the corresponding region was extracted from the visible light image using this frame.

$$\begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \times \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{2}$$

where x and y are the pixel coordinates of the visible light image; x_1 and y_1 are the coordinates of the pixels in the visible light image after registration; h_{11}(h_{22}) is the horizontal (vertical) scaling factor; h_{12} (h_{21}) is the horizontal (vertical) tilt factor; h_{13} (h_{23}) is the horizontal (vertical) translation factor; h_{31} and h_{32} are perspective transformation factors; h_{33} is the normalization factor, which is set to 1 in this paper; 1 in the matrix

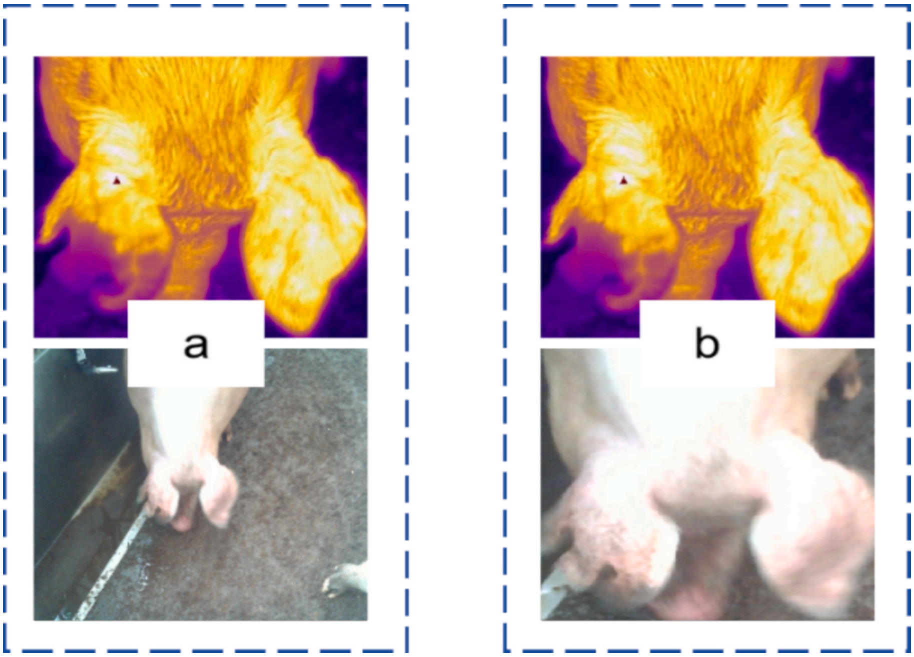

Fig. 4. Image registration effect. (a) Before registration. (b) After registration.

is used to introduce a point of the third dimension in the two-dimensional image transformation, so as to avoid the translation term in the image and reduce the computational complexity. The image registration effect is shown in Fig. 4.

3.2.2. Semantic segmentation

To automatically segment pig body contours and provide pixel-level key regions for temperature measurement from fused infrared and visible images, an improved DeepLabv3+ (Chen et al., 2018) semantic segmentation algorithm was developed to achieve both lightweight performance and high contour accuracy.

The enhanced DeepLabv3+ architecture achieves lightweight segmentation through three structural modifications: replacing the Xception backbone (Chollet, 2017) with MobileNetV4 (Qin et al., 2024), optimizing the dilation rates of atrous convolutions, and integrating CBAM (Woo et al., 2018) dual-attention modules at the encoder's output layer.

As illustrated in Fig. 5 (adapted based on (Chen et al., 2018)), the segmentation model consists of four main components: an input layer, an encoder, a decoder, and a prediction layer.

(1) Input layer

The input layer of the segmentation model takes as input a registered visible light image with a resolution of 512 × 384. To match the model's input requirements, the image is resized to 512 × 512.

(2) Encoder

In the encoder, the first deep convolutional neural network module (MobileNetV4) is used to perform feature extraction on the input pig image. Through the MobileNetV4 network, a high-level semantic feature map (32 × 32 × 320) and a low-level semantic feature map (128 × 128 × 24) are obtained. The low-level semantic features are passed directly to the decoder, while the high-level semantic features are forwarded to the Atrous Spatial Pyramid Pooling (ASPP) module (Chen et al., 2018). By applying dilated convolutions with varying dilation rates, the ASPP module further enhances feature extraction, enabling the capture of more discriminative information.

MobileNetV4 (Qin et al., 2024). achieves fast and accurate vision modeling for mobile and edge devices through an efficient lightweight design that combines depthwise separable convolution and pointwise convolution. In this study, depthwise separable convolution was adopted to reduce computational complexity. Specifically, the input pig image is decomposed into three channels (R, G, and B), after which a 3 × 3 convolution is independently applied to each channel, generating corresponding feature maps (Fig. 6).

Compared to traditional convolution, which uses a $K \times K$ kernel to process D_{in} channels, the computation required to obtain a feature map with D_{out} channels is $D_{\text{in}} \times D_{\text{out}} \times K \times K$. Notably, the computation is reduced to $D_{\text{in}} \times (D_{\text{out}} + K \times K)$ in this paper, which significantly decreases the computational cost and achieve a better lightweight model. At the same time, the 3 × 3 deep convolution with a step size of 2 has fewer parameters and computations. It reduces the amount of data processed by the subsequent layer, thus lowering the memory footprint and improving computational efficiency, as shown in Eq. (3).

$$Mobile_{MQA}(X) = Concat(attention_1, \cdots\cdots, attention_n)W^O$$

$$attention_j = Softmax\left(\frac{(XW^{Q_j})(SR(X)W^K)T}{\sqrt{d_k}}\right)(SR(X)W^V) \tag{3}$$

where SR represents the spatial downsampling performed by the deep convolution module with step size of 2, W is the weight matrix, $Softmax$ is an activation function that maps the input value to the probability distribution between 0 and 1; d_k is the dimension of the input vector; and Q_j is the j-th transformed vector of the input.

The ASPP module consists of one standard convolution, three dilated convolutions with dilation rates of 12, 24, and 36, and one pooling layer, all operating in parallel. To minimize the feature loss, the outputs from these five parallel operations are combined and fused using the Concat module. Subsequently, the number of channels is adjusted through a 1 × 1 convolution, after which the CBAM module is applied to enhance feature representation by integrating both channel and spatial attention, thereby improving the learning of pig-specific features.

CBAM(Woo et al., 2018) is a lightweight dual-attention mechanism that integrates a Channel Attention Module (CAM) and a Spatial Attention Module (SAM). It enhances the extraction of key features from the pig's body by refining information at both the channel and spatial levels (Fig. 7 adapted from (Woo et al., 2018)). When combined with the ASPP module, CBAM further strengthens feature representation, ensuring that both multi-scale context and attention-guided details are effectively captured.

The CAM evaluates the relative importance of different feature channels and assigns corresponding weights, enabling the segmentation model to focus more effectively on channels that are critical for RoIs on the pig's body surface relevant to temperature detection. The SAM

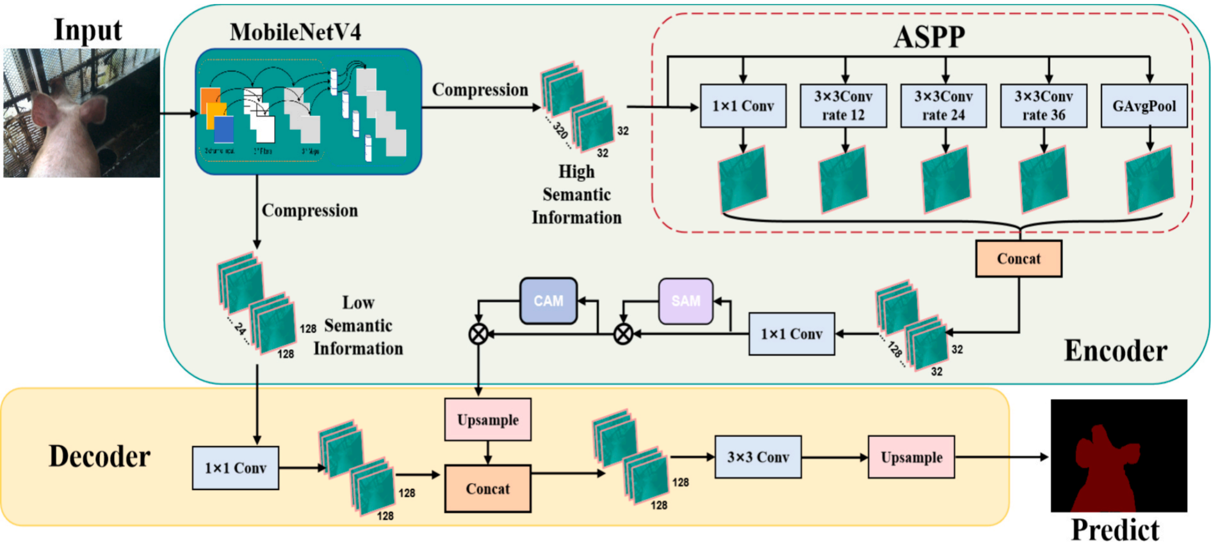

Fig. 5. Structure of the semantic segmentation model.

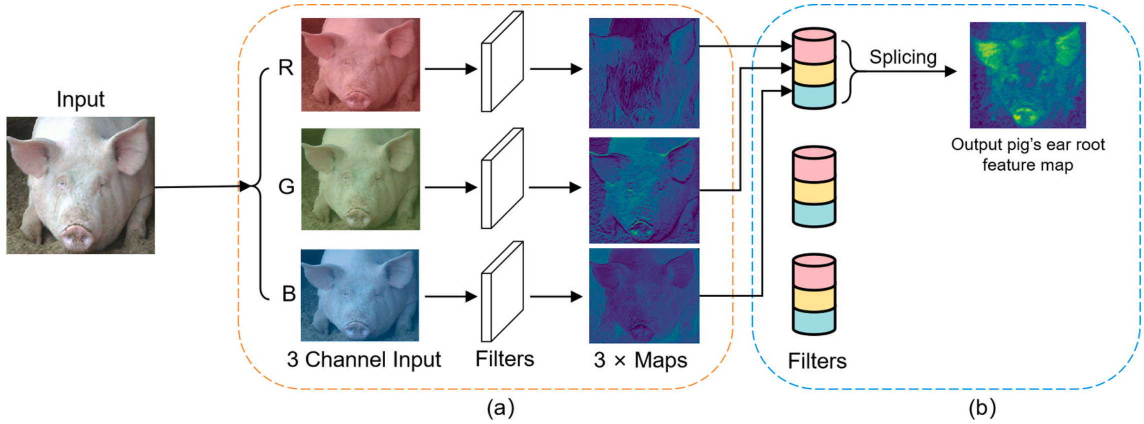

Fig. 6. Depth-separable convolution. (a) Depthwise Convolution. (b) Pointwise Convolution.

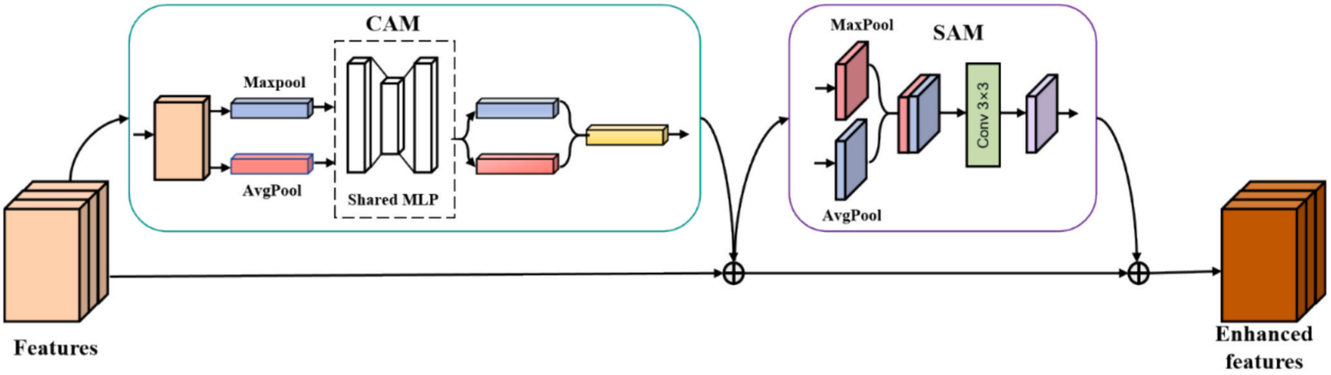

Fig. 7. Structure of CBAM.

operates in the spatial domain, enhancing the recognition of contour features on the pig's body surface and improving segmentation accuracy.

(3) Decoder

The number of low-level semantic channels is first adjusted using a 1 × 1 convolution to match the channel dimensions of the decoder output. The resulting features are then further fused with a 3 × 3 convolution. After the final upsampling and processing step, all aggregated features are passed into the prediction layer for output generation.

(4) Output layer

The size of the upsampled feature map is adjusted so that the prediction output matches the resolution of the input image. Each pixel is then classified using the Softmax function, enabling automatic segmentation of the pig's body.

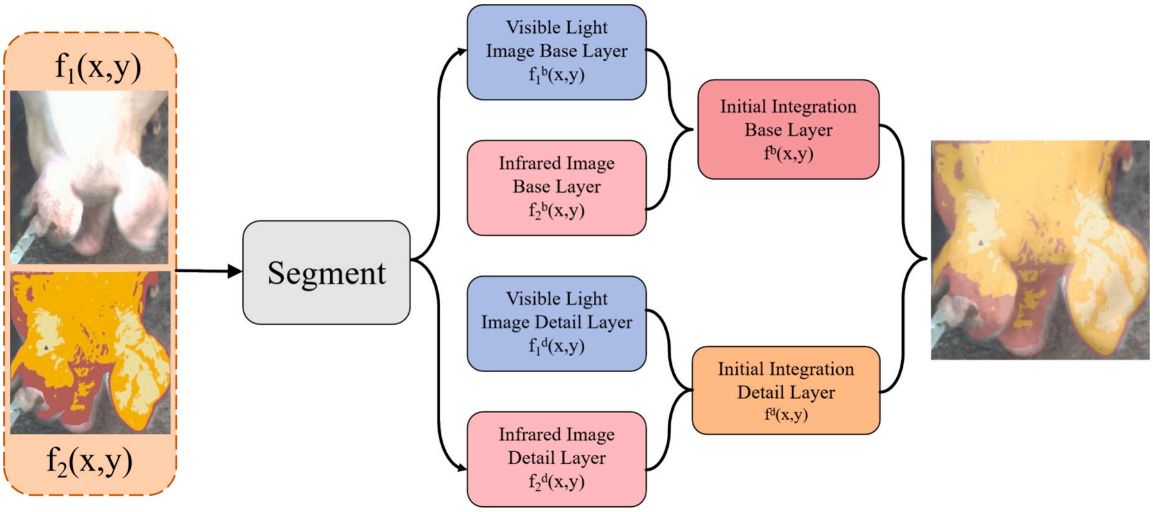
f1(x,y)
f2(x,y)
Segment
Visible Light Image Base Layer f1b(x,y)
Infrared Image Base Layer f2b(x,y)
Visible Light Image Detail Layer f1d(x,y)
Infrared Image Detail Layer f2d(x,y)
Initial Integration Base Layer fb(x,y)
Initial Integration Detail Layer fd(x,y)

Fig. 8. Flow chart of infrared and visible light image fusion.

3.3. Image fusion

Since infrared thermal images contain skin temperature information but have lower resolution and fewer details, while visible light images provide richer detail for accurate identification, image fusion was applied to enable precise temperature extraction (Fig. 8). The module takes as input a visible light image and its corresponding infrared image. Both images are decomposed into scale layers: the visible light base layer $f_1^b(x,y)$, the visible light detail layer $f_1^d(x,y)$, the infrared base layer $f_2^b(x,y)$, and the infrared detail layer $f_2^d(x,y)$. Different fusion strategies are then applied to the base and detail layers, respectively (Fig. 8).

(1) The base layer and detail layer of the visible light image (f_1^b (X, Y) and $f_1^d(X, Y)$) and the base layer and detail layer of the infrared thermal infrared image (f_2^b (X, Y) and $f_2^d(X, Y)$) are obtained after decomposition using a mean filter (window size is 35).

(2) The base layer is obtained using the weighted average method to calculate the base layer $f^b(x,y)$ after the preliminary fusion. The detail layer is derived by calculating the Euler distance to obtain the fusion coefficient matrix $\xi_1(x,y)$ *and* $\xi_2(x,y)$. The preliminary fusion of the detail layer $f^d(x,y)$ is then obtained, as shown in Eq. (4).

$$f^d(x,y) = \varepsilon_1(x,y)f_1^d(x,y) + f_2^d(x,y)\varepsilon_2(x,y) \tag{4}$$

3.4. Temperature extraction

3.4.1. Detection model for the RoI on pig body surface

The body surface key temperature identification model is composed of five main components: Input, Backbone, Neck, Head and Output (Fig. 9).

(1) Input layer

The input layer consists of fused images that have been pre-processed and normalized to 640 × 640 pixels. To standardize the inputs, partially registered images are resized to 640 × 640, ensuring stability and consistency in model processing.

(2) Backbone

The backbone network is primarily responsible for multi-scale feature extraction from pig body surface images to ensure accurate detection. A parallel backbone architecture was introduced to capture contextual information through multi-path computation units, where Vision Mamba encodes global features via parallel self-attention modules to establish cross-region correlations. At the same time, the network extracts local textures through nonlinear transformations while integrating global information, thereby achieving comprehensive surface sensing.

Regional Feature Discriminative Adaptive Processor (RFDAP): The RFDAP further refines this process and consists of three components (Fig. 10): frequency feature construction, spatial feature adaptation, and band feature modulation.

Frequency Feature Construction (FFC): The workflow for frequency feature construction is as follows: First, the input image data $X \in \mathbb{R}^{H\times W\times C_{in}}$ is equally divided by channel dimension into n channel grouping data blocks $X \in \mathbb{R}^{H\times W\times \frac{C_{in}}{n}}(i = 1, 2, \cdots n)$, each channel-grouped data block X_i carries information of a specific image channel. Then, X_i is transformed from the spatial domain to the frequency domain by Discrete Fourier Transform (DFT) to obtain the frequency domain feature representation $F(X_i)$. DFT and the inverse DFT (iDFT) are calculated as shown in Eqs. (5) and (6).

$$F(k,l) = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n) \times e^{-j2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right)} \tag{5}$$

$$f(m,n) = \sum_{k=0}^{M-1}\sum_{l=0}^{N-1} F(k,l) \times e^{-j2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right)} \tag{6}$$

where $f\ (m, n)$ is the discrete signal in the spatial domain (e.g., image pixel values); $F\ (k, l)$ is the discrete spectrum in the frequency domain; M *and* N are the height and width of the signal; j is an imaginary unit; and $e^{j2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right)}$ is the frequency-domain basis function characterizing the phase and amplitude of the frequency components.

Second, the frequency domain data $F(X_i)$ is sorted by frequency and processed through the Fully-Connected (FC) layer to generate the

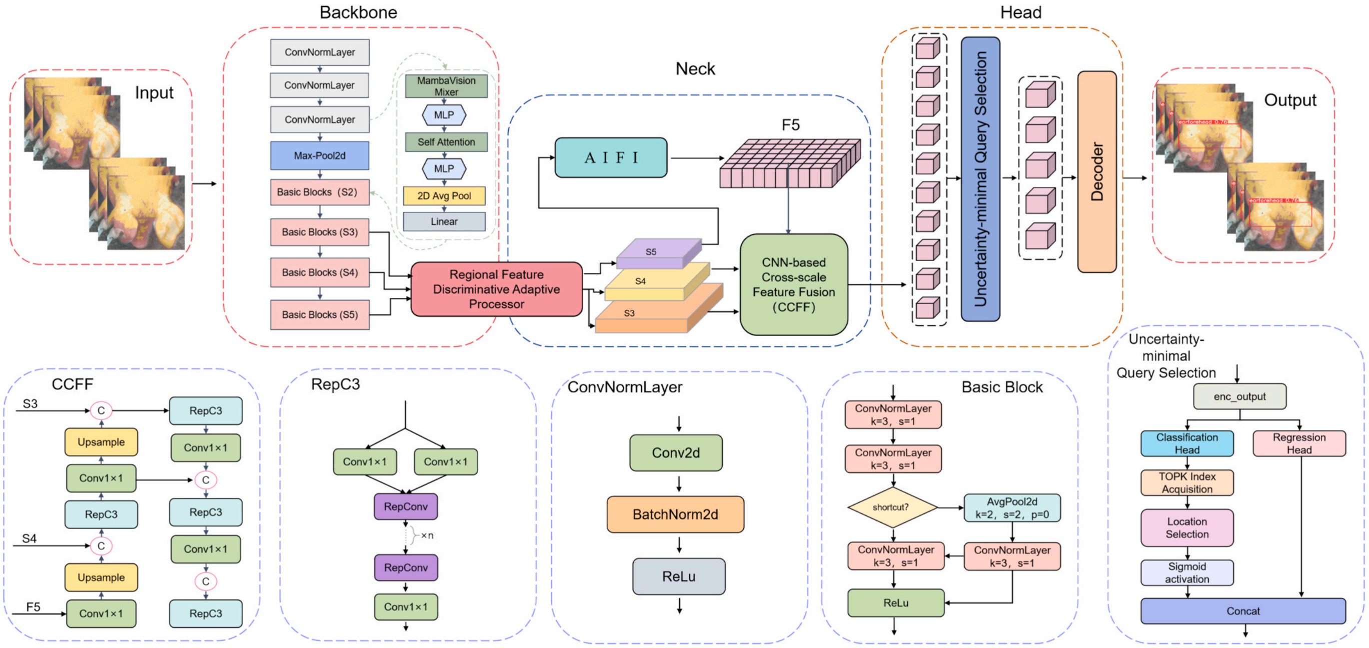

Fig. 9. Structure of the detection model for the RoI on pig body surface.

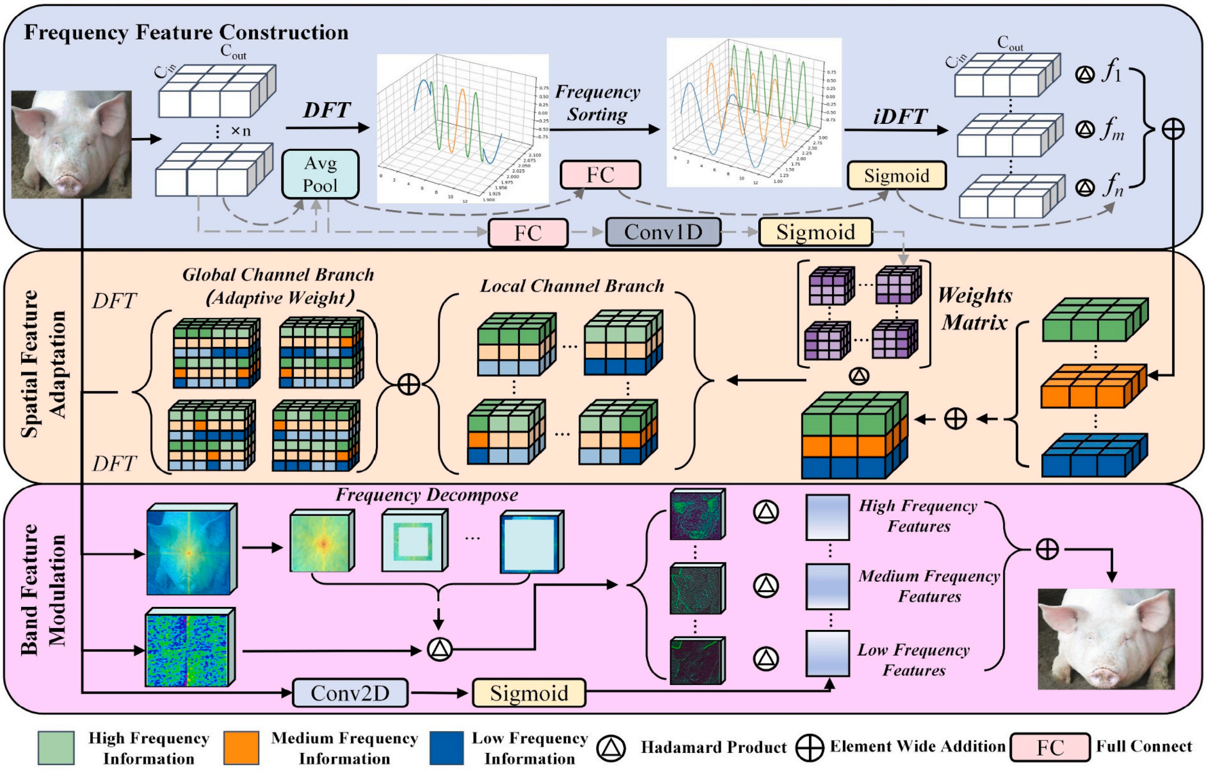

Fig. 10. Regional feature discriminative adaptive processor.

feature representation $FC(sort(F(X_i)))$, and then processed by the Sigmoid function to obtain the modulation factor $f_i = \sigma(FC(sort(F(X_i))))$.

The frequency domain data is transferred back to the spatial domain by the inverse iDFT to obtain W_i ($W_i = F^{-1}(F(X_i))$). Eventually, the group W_i are summarized by Hadamard product with the corresponding modulation coefficients f_i, and the output feature $F_1 = \sum_{i=1}^{n} f_i \cdot W_i$ s obtained. The Hadamard product formula is shown in Eq. (7).

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \times \begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \cdots & b_{nn} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} & \cdots & a_{1n}b_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1}b_{n1} & \cdots & a_{nn}b_{nn} \end{pmatrix} \tag{7}$$

Spatial Feature Adaptation (SFA): The spatial feature adaptation process (Fig. 11 (adapted from (Chen et al., 2025))) is designed to strengthen FP-DETR's capacity to capture features from the input data by accurately modulating weights, thereby improving the model's adaptability and representational power.

In this study, a 1-D convolution is applied to the local channel to effectively capture local channel information while significantly reducing computational cost. To address the limited utilization of global feature information in local branches, additional global channels are introduced to aggregate global context, followed by predicting a modulation value across the input channel, output channel, and kernel dimensions.

Band Feature Modulation (BFM): Although the weights generated by the Frequency Feature Construction (FFC) and Spatial Feature Adaptation (SFA) modules enhance representation, they still maintain spatial invariance at the global level. To address this limitation, BFM is incorporated to achieve targeted weight conversion for different frequency components. Specifically, feature frequency decomposition is performed by applying a frequency-domain transform to the input feature map X,

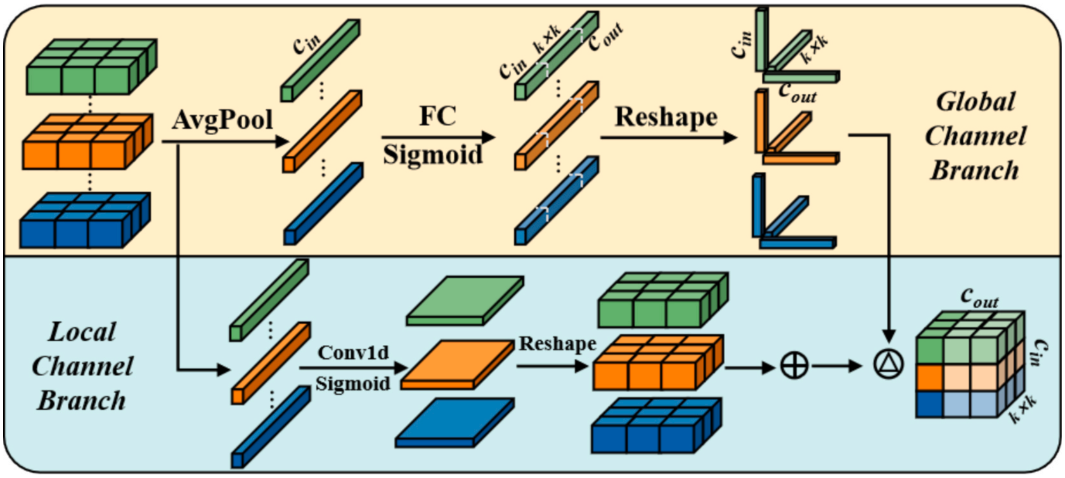

Fig. 11. Structure of spatial feature adaptation.

and binary masks M_b (with $b = 1, ..., B$, and $B = 3$) are used to separate it into different frequency bands X_b ($X_b = F^{-1}(M_b \odot F(X))$), where F and F^{-1} are the DFT and the iDFT); then the frequency feature modulation is performed, and the spatially variable modulation coefficients $A_b(A_b = Sigmoid(Conv\,2\,D(X_b)))$. Each frequency band is then modulated by generating spatially varying coefficients A_b ($A_b = Sigmoid(Conv\,2\,D(X_b)))$ through a convolutional layer followed by a Sigmoid activation. Finally, the modulated frequency bands are fused through element-wise summation to obtain the output feature Y $\left(Y = \sum_{b=0}^{B-1}(A_b \odot X_b)\right)$, enabling FP-DETR to adaptively capture the complex spatial–frequency structure of the image and improve the representation of both high- and low-frequency features.

(3) Neck

The Adaptive Interaction Feature Integration (AIFI) module is designed to further enhance feature representations by fully leveraging global information (Fig. 12). The input feature map has dimensions (B, C, H, W), where B denotes the batch size, C the number of channels, and H and W the spatial height and width, respectively.

First, the Conv1-D and Embedding modules are applied to the input feature map, reducing the number of channels by half and producing a feature map with dimensions (B, $C/2$, H, W). Next, the spatial dimensions H and W are flattened and reshaped into (B, $C/2$, $H \times W$) to prepare for subsequent processing. Positional information is then incorporated using sine–cosine embedding, which enables FP-DETR to accurately capture the spatial positions of features. The calculation process is shown in Eqs. (8) and (9).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{8}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right) \tag{9}$$

where PE is the positional encoding matrix used to give positional information to the elements of the model sequence; pos means the index of the element's position in the sequence (from 0), i is the index of the encoding vector dimension (also from 0, $2i$, $(2i + 1)$ corresponds to the even and odd dimensions, respectively); d_{model} refers to the hidden layer dimension of the model; *10,000* is a fixed constant used to scale the exponential function and modulate the variation of the encoding period.

After that, the Multi-Head Attention module is applied to capture long-range dependencies and contextual relationships among features, followed by a Feed Forward Neural Network (FFN) that further refines the feature representation through nonlinear transformations. Finally, the enhanced features are projected back to the input dimensions using a Linear layer and classified with the Softmax function.

(4) Head

The input data are first processed by the Uncertainty-Minimum Query Selection Module (Zhao et al., 2024), which filters out non-essential information. The refined data are then concatenated and activated to determine the TOP-K indices for location selection.

(5) Output

The output layer is responsible for predicting both the target's position (bounding box) and its classification (confidence score) based on the results from the Head layer. The category score map (*Sclass*) is generated according to Eq. (10). Regression is then applied to produce the bounding box parameters, including the center coordinates (x_c, y_c), confidence score p, width w and height h as defined in Eqs. (11)–(15).

$$\begin{gathered} S_{class}(x,y,c) = Softmax(f_1(x) \times f_2(x) + b_{class(c)}) \\ f_1(x) = \sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{k}\sum\nolimits_{d=1}^{k} w_{class}(i,j,d,c) \end{gathered} \tag{10}$$

$$\begin{gathered} f_2(x) = F_{Neck}(x+i,y+j,d) \\ x_c = \sigma\left(\sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{k}\sum\nolimits_{d=1}^{C_{Neck}} w_x(i,j,d) \times F_{Neck}(x+i,y+j,d) + b_x\right) \end{gathered} \tag{11}$$

$$y_c = \sigma\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{d=1}^{C_{Neck}} w_y(i,j,d) \times F_{Neck}(x+i,y+j,d) + b_y\right) \tag{12}$$

$$p = \sigma\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{d=1}^{C_{Neck}} w_p(i,j,d) \times F_{Neck}(x+i,y+j,d) + b_p\right) \tag{13}$$

$$w = e^{\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{d=1}^{C_{Neck}} w_c(i,j,d) \times F_{Neck}(x+i,y+j,d) + b_w} \tag{14}$$

$$h = e^{\sum_{i=1}^{k}\sum_{j=1}^{k}\sum_{d=1}^{C_{Neck}} w_h(i,j,d) \times F_{Neck}(x+i,y+j,d) + b_h} \tag{15}$$

where σ represents Sigmoid activation function to limit the output between 0 and 1; and e^x is used to keep width and height at positive values.

Finally, the model outputs the coordinates (x_c, y_c, w, h) of the bounding box and the confidence score p for each object in the image.

3.4.2. Temperature extraction

The temperature extraction is divided into three parts: (1) the fused image is fed into the RoI detection module to obtain the horizontal and vertical coordinates of the bounding box; (2) all temperature values within the defined region are extracted from the fused image; (3) the

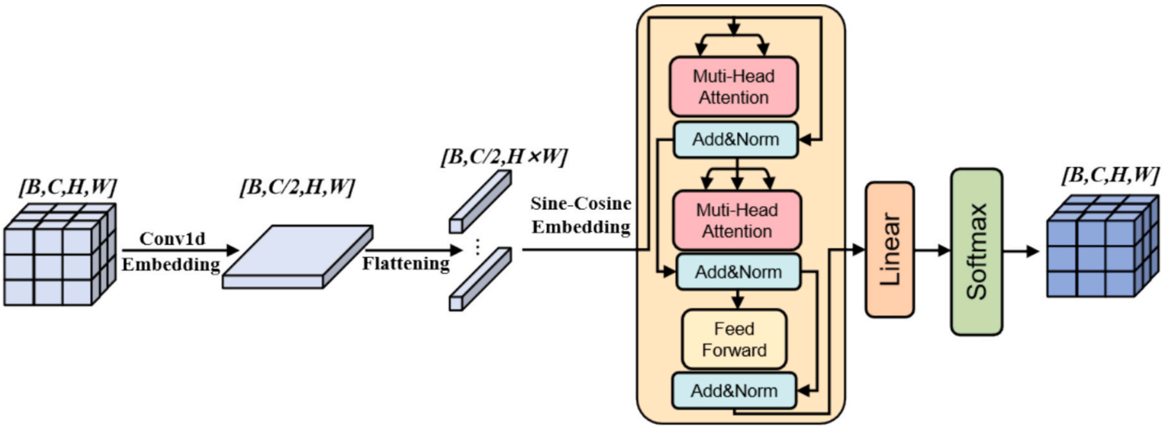

Fig. 12. Structure of AIFI, which integrates a convolution model with multi-head attention from a typical transformer model.

maximum temperature within the RoI is selected to represent the pig's body surface temperature.

3.5. Model evaluation indicators

Precision (P), Recall (R), FPS, mAP@50, F1 Score and calculations volume of Parameters are used to evaluate the detection model performances as shown in Eqs. (16)–(21), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-Square (R^2), Confidence Interval (CI) (95%) and t (p-value) are used to evaluate the accuracy of temperature extraction as shown in Eqs. (22)–(26).

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{16}$$

$$\mathrm{R} = \frac{TP}{TP - FN} \times 100\% \tag{17}$$

$$\mathrm{FPS} = \frac{1}{p_t} \tag{18}$$

$$\mathrm{Confidence} = P_r(Object) \times IoU_{pred}^{truth} \tag{19}$$

$$\mathrm{F1Score} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{20}$$

$$\mathrm{mAP@50} = \frac{\sum_{i=1}^{C} AP_i}{C} \tag{21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{22}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \widehat{y}_i)| \tag{23}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{24}$$

$$t = \frac{\overline{d}}{\frac{s_d}{\sqrt{n}}} \tag{25}$$

$$\mathrm{CI} = \left(\overline{d} - t_{\frac{\alpha}{2}} \times \frac{s_d}{\sqrt{n}}, \overline{d} + t_{\frac{\alpha}{2}} \times \frac{s_d}{\sqrt{n}},\right) \tag{26}$$

where *TP* is the number of targets in the key areas that are correctly detected in the pig image; *FP* is the number of targets that identify errors in critical areas; *FN* is the number of targets that are identified as false but are true; y_i and $\widehat{y}_i$ are the *i*-th true and predicted values; *n* is the number of image test sets; $\overline{y}$ is the sample mean; p_t is an abbreviation for *Processing time per frame*, which is the time it takes for the model to detect each frame of image; *P_r(Object)* is the probability of the existence of the object in the bounding box, if there is an object, *P_r(Object)* = 1, otherwise *P_r(Object)* = 0; *IoU* is the intersection ratio of the real box (ground truth) and the predicted box (predicted *IoU* is the intersection ratio of real box (ground truth) and predicted box (predicted box); *C* is the total number of categories; AP_i represents the AP value of the *i*-th category; $\overline{d}$ is the mean of the differences; s_d is the standard deviation of the differences; $t_{\alpha/2}$ is is the *t*-distribution critical value for a given significance level α.

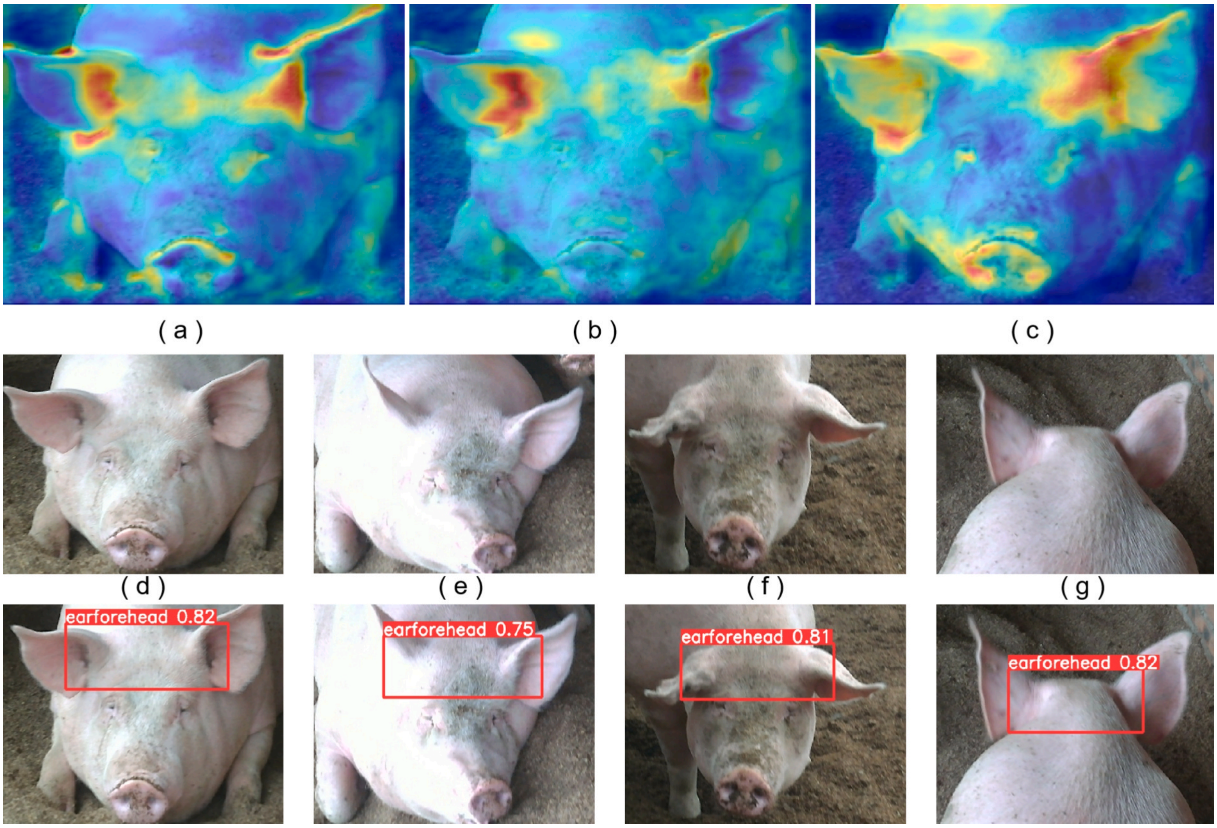

Fig. 13. Visualization of the detection and segmentation model. (a) Original Backbone. (b) Vision-Mamba Backbone. (c) Parallel backbone. (d) Normal. (e) Bright light. (f) Face masking. (g) Top view.

4. Results and discussion

4.1. Performance analysis of the detection and segmentation model

This part systematically illustrates the differences in feature extraction across architectures through visual analysis of feature heatmaps (Fig. 13(a)–(c)). The traditional backbone (a) exhibits localized receptive field characteristics, with activations concentrated on specific regions of the face, such as the ear and forehead, while its capacity to model cross-regional correlations is markedly limited. By contrast, the Vision Mamba architecture (b) presents a more globally distributed activation pattern, indicating its ability to establish long-range dependencies across regions through a bidirectional state space modeling mechanism. Its responses not only extend over a broader spatial range but also display superior spatial continuity, highlighting the advantages of state space models in capturing global contextual information.

The proposed parallel hybrid architecture (c) integrates the strengths of both approaches. Compared with a single traditional backbone, it preserves local feature extraction while substantially broadening the spatial coverage of activations; compared with Vision Mamba alone, it enhances activation intensity in critical regions. Traditional backbones excel at capturing fine-grained local details through layered receptive fields, whereas Vision Mamba establishes semantic associations across distant regions through global modeling. Their combination produces complementary effects across multiple scales, yielding feature representations that remain sensitive to local detail while providing global contextual understanding, which aligns with the expected outcomes of the experiment.

The detection results for Fig. 13(d–g) demonstrate FP-DETR ’s performance under various conditions, with detailed analysis provided for (d) normal light, (e) bright light, (f) face masking, and (g) top view scenarios.

In the normal light condition (Fig. 13(d)), FP-DETR successfully identifies the RoI with a high confidence of 0.91. This result benefits from uniform lighting, clear image quality, and well-defined facial features, such as contours and textures, which closely match the training data. Under such conditions, the model can achieve stable recognition with high confidence.

In the bright light condition (Fig. 13(e)), the confidence decreases to 0.75, though it still remains reliable. The main reason for this reduction is local overexposure and reflections caused by strong illumination, which obscure critical features such as the eyes and mouth. Furthermore, changes in lighting alter the image’s color distribution, making it harder for FP-DETR to extract sufficient discriminative features, thus

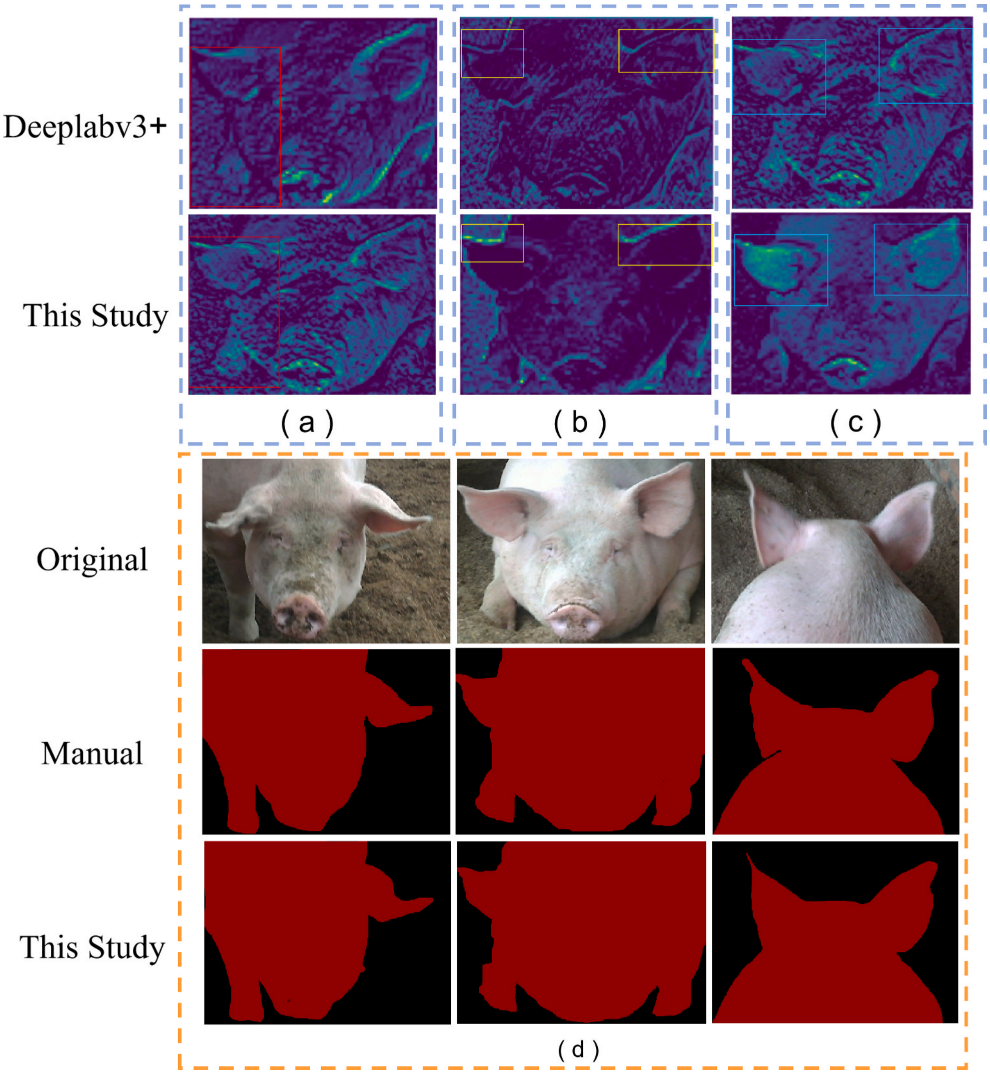

Fig. 14. Visualization of the segmentation model. (a) Foreground. (b) Background. (c) Segmentation result. (d) Comparison of manual and model. In the upper dashed box, (a), (b), and (c) correspond to the 1st, 2nd, and 3rd column figures (foreground, background, and the result) respectively; (d) in the lower orange box corresponds to all figures in this area, showing the three-group comparison (original image, manual segment and segmentation result of the proposed method) for pigs from different views.

lowering the confidence slightly.

For the face masking condition (Fig. 13(f)), the model reaches a confidence of 0.81. Occlusion by dirt or padding directly covers some facial features, leaving FP-DETR to rely on visible regions like the eyes and forehead. While this increases recognition uncertainty, the combination of local and global cues still allows for reasonably accurate predictions, though with lower confidence than in normal conditions.

In the top-view condition (Fig. 13(g)), FP-DETR achieves a confidence of 0.82. Changes in perspective modify the geometric structure of the face, causing some key features to become distorted or partially invisible. This creates abnormal proportions that complicate detection. Nevertheless, by leveraging stable features such as ear positions and their relative alignment with the body, FP-DETR adapts effectively to these spatial changes, maintaining robust performance across varying viewpoints.

The effects of the improved DeepLabv3+ model on foreground (a), background (b), and final segmentation results (c) are presented in Fig. 14(a–c). The highlighted regions represent the key feature areas identified by FP-DETR.

For foreground segmentation (Fig. 14(a)), the improved model shows a clear increase in highlighted regions, such as the more distinct left front hoof, capturing a greater number of feature points. This enhancement allows more accurate localization of pig body contours and internal feature details, thereby improving recognition accuracy.

In background segmentation (Fig. 14(b)), the model demonstrates better contour extraction, successfully identifying fine structures such as the ears and front paws. This improves detailed feature capture and contributes to more accurate body contour delineation.

For the final segmentation results (Fig. 14(c)), the pig's overall body contour appears smoother, with highlighted areas concentrated on the body region, reducing background misclassification. These results confirm that FP-DETR significantly improves the extraction and representation of pig body surface features, ensuring more precise segmentation for downstream temperature detection.

Across the three test samples (Fig. 14(d)), the segmentation results of the proposed model show high consistency with manual annotations. The calculated Intersection over Union (IoU) values were 0.914, 0.920, and 0.849, with an average of 0.894. This indicates that nearly 90% of the segmented regions overlap with expert annotations, demonstrating strong reliability. The best performance (IoU = 0.920) was observed in relatively simple contours, while the lowest value (IoU = 0.849) occurred in more complex structures with sharp edges. A closer inspection reveals that discrepancies mainly appear along fine boundaries, such as the tips of ears and abrupt contour transitions, where manual annotations capture subtle edges with higher precision. Despite these minor differences, the model maintains smooth and coherent contours, achieving segmentation quality that is comparable to human annotation and sufficient for reliable temperature extraction tasks.

In summary, FP-DETR shows stable recognition performance under different lighting conditions (normal, strong light), facial occlusion, and top-down view angles. Under normal lighting (Fig. 13(d)), FP-DETR achieves a confidence level of 0.91, due to uniform lighting and clear facial features; under strong light (Fig. 13(e)) and occlusion (Fig. 13(f)) conditions, the confidence level remains above 0.75, indicating that FP-DETR has a certain degree of robustness to local feature loss and lighting interference. The confidence level of 0.82 under a top view (Fig. 13(g)) further verifies FP-DETR 's adaptability to different angle changes.

The improved Deeplabv3+ model performs exceptionally well in image segmentation tasks: foreground segmentation (Fig. 14(a)) significantly improves the ability to capture key feature points (such as the left front hoof); background segmentation (Fig. 14(b)) optimizes contour recognition accuracy (such as ears and front paws); the final result (Fig. 14(c)) presents a smoother pig body contour segmentation with reduced false detection rates. Experiments have proven that FP-DETR effectively improves feature extraction accuracy and robustness in complex scenes by integrating frequency domain and spatial domain information.

4.2. Comparation with other models

In testing the models, this study conducted five repeated experiments, where each training was performed under exactly the same hardware and software environment. The final results were subjected to significance testing, and the results are shown in Table 2.

As shown in Table 2, the FP-DETR model achieves 99.07% precision and 96.75 ± 0.48% recall with (5.9 ± 0.07) × 10^6 parameters and 8.79 ± 0.28 hours training time. Compared with Faster R-CNN and YOLOv8, FP-DETR achieves an overall improvement in parameters, FPS, and accuracy, with a 91.8% and 7.37% reduction in parameters, a 50 FPS and 34 FPS increase in frame rate, and a 9.63% and 4.29% increase in precision.

Compared with YOLOv9, YOLOv12, and RT-DETR, despite the parameters increase of 11.72%, 1.03%, and 39.72%, the frame rate is significantly improved by 40 FPS, 38 FPS, and 26 FPS, and the precision is improved by 6.42%, 5.23% and 4.95%, and the FP-DETR has significantly improved the detection efficiency and accuracy while considering the number of parameters, showing excellent comprehensive performance.

In terms of storage size, FP-DETR requires 16.9 MB, which is slightly larger than YOLOv12 (15.8 MB) and comparable to YOLOv9 (17.2 MB), while being much smaller than Faster R-CNN (100.2 MB). This compact model size, combined with high accuracy and speed, makes FP-DETR suitable for deployment on edge devices with limited hardware resources.

Compared with the recently proposed YOLOv12 model, FP-DETR demonstrates superior performance in both accuracy and efficiency. Specifically, FP-DETR achieves a higher precision (+5.22%) and recall, while delivering a significantly faster frame rate (+38 FPS). Although the number of parameters is slightly higher (+1.2%), the lightweight dual-backbone design and the integration of frequency–spatial domain fusion allow the model to better suppress environmental noise and enhance feature extraction. This balance of accuracy, speed, and parameter efficiency highlights the advantages of FP-DETR in real-time pig body temperature detection, making it more suitable for practical deployment under large-scale farm conditions.

Table 3 presents the P-value calculations for FP-DETR compared to the baseline models. These P-values help further validate the superiority of FP-DETR over the baseline models in terms of precision, recall, and FPS. In all comparisons, the P-values for precision, recall, and FPS are consistently below the 0.05 threshold, indicating that the observed improvements are not due to random variation but are statistically robust and reliable.

For instance, FP-DETR achieved markedly higher precision than YOLOv9, YOLOv12, and RT-DETR, with P-values of 5.92 × 10^{-7}, 8.35 × 10^{-7}, and 3.10 × 10^{-6}, respectively. Similarly, recall improvements over YOLOv8 and YOLOv12 yielded P-values of 2.55 × 10^{-6} and 9.87 × 10^{-6}, while FPS comparisons demonstrated significant advantages even against RT-DETR (P = 9.55 × 10−4). These results indicate that FP-DETR's enhancements in accuracy, recall, and processing speed are statistically significant and reproducible, thereby reinforcing the reliability and scientific rigor of the proposed model.

Therefore, the FP-DETR model can achieve good detection accuracy and is a lightweight model, it has possibility to be deployed on edge devices with limited hardware performance.

4.3. Comparisons to the existing detection methods

There are some previous studies on the automatic detection for the RoI on pig body surface, the comparison results are shown in Table 4. For example, Guo, et al. (2023) developed a deep learning framework for individual pig detection and tracking, achieving a detection accuracy of 94.72%, 7.5 × 10^6 parameters and a speed of 12 FPS. In our research

Table 2
Comparison with different models.

Model	Precision (%)	Recall (%)	Size (MB)	Parameters/10^6	TrainTime (h)	FPS (frame/s)
Faster R-CNN	89.44 ± 0.49	87.60 ± 0.58	99.30 ± 0.29	72.06 ± 0.06	9.84 ± 0.30	17 ± 5
YOLOv8	94.78 ± 0.40	93.15 ± 0.57	16.43 ± 0.27	6.38 ± 0.04	10.14 ± 0.21	33 ± 4
YOLOv9	92.65 ± 0.52	91.95 ± 0.39	17.20 ± 0.39	5.29 ± 0.03	11.66 ± 0.21	27 ± 2
YOLOv12	93.84 ± 0.53	92.62 ± 0.38	15.87 ± 0.20	5.85 ± 0.06	9.23 ± 0.36	29 ± 6
RT-DETR	94.12 ± 0.53	93.68 ± 0.61	15.20 ± 0.34	4.23 ± 0.05	10.29 ± 0.19	41 ± 6
This study	99.07 ± 0.50	96.75 ± 0.48	16.79 ± 0.21	5.91 ± 0.07	8.79 ± 0.28	67 ± 10

Table 3
P-value calculations of FP-DETR compared with baseline models.

Model (FP-DETR)	P value		
	Precision	Recall	FPS
Faster R-CNN	1.60×10^{-1o}	3.42×10^{-9}	4.91×10^{-6}
YOLOv8	2.48×10^{-8}	2.55×10^{-6}	2.86×10^{-4}
YOLOv9	5.92×10^{-7}	1.13×10^{-5}	7.44×10^{-4}
YOLOv12	8.35×10^{-7}	9.87×10^{-6}	1.21×10^{-3}
RT-DETR	3.10×10^{-6}	6.42×10^{-5}	9.55×10^{-4}

Table 4
Comparison with existing methods.

Literatures	Specific algorithms	Precision (%)	FPS (frame/s)	Parameters (10^6)
(Guo, et al., 2023)	CNN	94.72	12	7.5×10^6
(Xie, et al., 2023)	YOLOv5s-BiFPN model	96.88	100	5.3×10^6
(Zhang, et al., 2024)	YOLOv7-tiny-Ghost	93.5	10	3.59×10^6
This study	FP-DETR	98.9	68	5.9×10^6

group, Xie, et al. (2023) developed an infrared thermal imaging detection method based on YOLOv5s-BiFPN with detection accuracy of 96.88%, the frame rate of 100 FPS, and the number of parameters of 5.3 × 10^6. Zhang, et al. (2024) proposed a temperature extraction algorithm based on registered images, although with less parameters of 3.59 × 10^6, the precision of 93.5% is relatively low.

The accuracy of FP-DETR is 98.9% (+4.18% compared to Guo, +2.02% compared to Xie, +5.4% compared to Zhang), although the FPS (68) is lower than that of YOLOv5s-BiFPN (100), it is over the 60-frame limit of the edge device, and 6.8 times faster than that of YOLOv7-tiny-Ghost (10FPS). The number of parameters (5.9 M) is only slightly higher than that of YOLOv5s (5.3 M) and much lower than that of CNN (7.5 M), which achieves the best balance of accuracy-speed-lightweight.

4.4. *Automatic temperature extraction and verification*

To assess the accuracy of FP-DETR in extracting surface temperatures from key body regions of pigs, the automatically extracted values were compared with manually measured ones (Fig. 15). The results show a strong correlation, with an R^2 of 0.957, while the MAE and RMSE were 0.108 and 0.142, respectively. These metrics indicate that FP-DETR provides highly consistent temperature estimates, closely matching manual measurements and ensuring reliable application in practical settings.

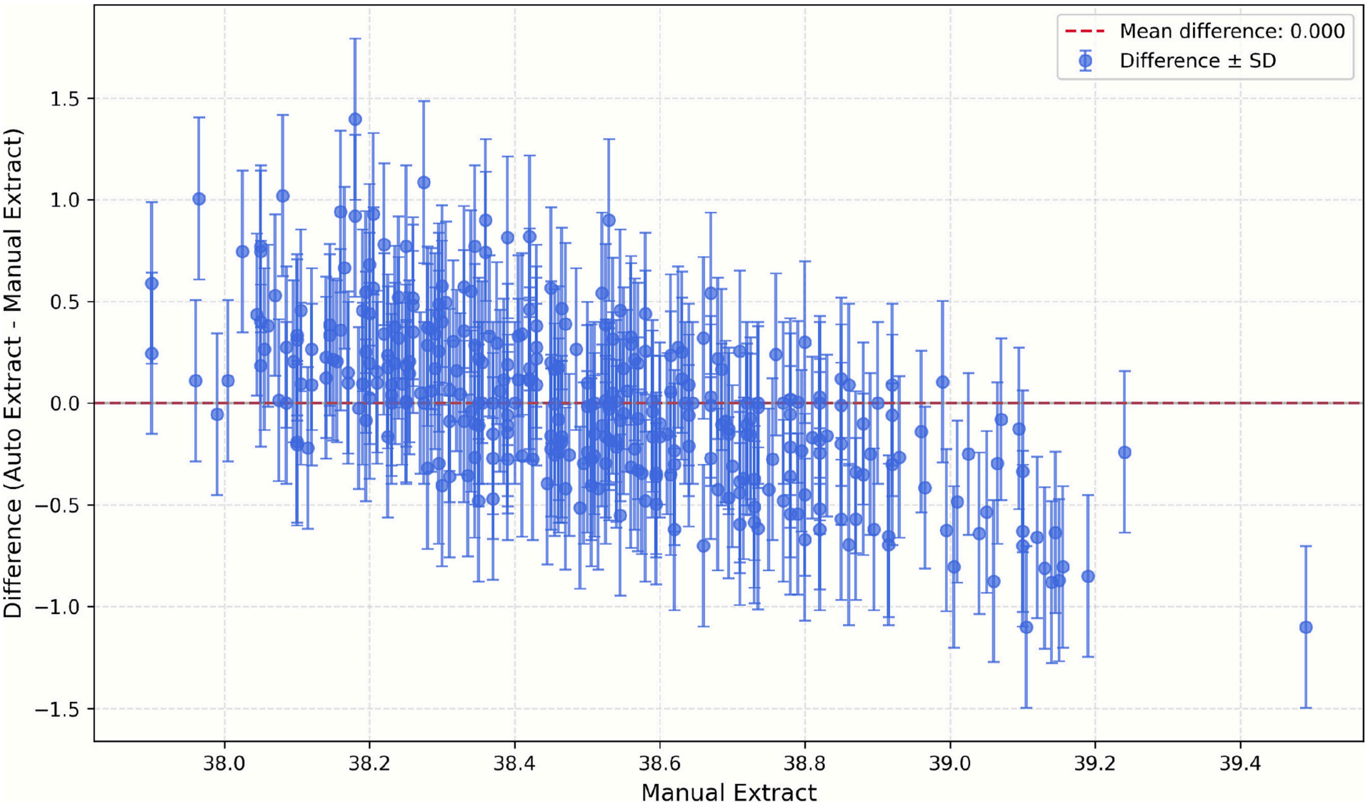

Fig. 15. Fitting effect of model and manual extraction temperature.

Fig. 15 illustrates the differences between manually extracted and model-extracted temperatures across various ranges. On the plot, the horizontal axis represents manual extraction values, while the vertical axis shows the difference between automatic and manual extraction (Auto Extract – Manual Extract). The mean difference is close to zero, indicating strong consistency between the two methods and validating FP-DETR's overall accuracy. Within the 38.0–38.6°C range—the normal temperature interval for pigs—data points are densely clustered and align closely with the zero line. This pattern demonstrates that model-based extraction matches manual measurements particularly well in this physiological range, confirming FP-DETR's high precision and strong applicability under typical farm conditions.

What's more, the statistical analysis reveals that the automatic extraction algorithm we proposed shows a very small difference compared to manual extraction. The mean difference between the two methods is only 0.0009, indicating that the temperatures extracted by both methods are nearly identical. The 95% confidence interval for the difference is between −0.0414 and 0.0431, suggesting that the true difference is negligible and close to zero.

Furthermore, the p-value from the paired *t*-test is 0.9683, which is significantly greater than the 0.05 threshold, indicating that there is no statistically significant difference between the two methods. This implies that the observed differences are likely due to random variation rather than a systematic bias.

4.5. *Ablation experiments*

4.5.1. *Model ablation experiment*

As shown in Table 5, compared with the baseline, the Dual-Backbone combined model improves precision and F1 score by 0.91% and 1.05%, respectively, although mAP@50 decreases by 0.82%. Multi-scale and multi-level extraction offered by the dual-backbone design enriches feature representation, supporting more accurate recognition and consistent classification. At the same time, however, the interaction between features from both backbones may disrupt certain high-recall characteristics, thereby reducing average accuracy at the 50% IoU threshold. The RT-DETR with the RFDAP module yields improvements of 1.14% in precision, 1.09% in F1 score, and 0.28% in mAP@50. These gains result from the RFDAP module's ability to strengthen attention mechanisms, increase feature sensitivity, and optimize feature quality, allowing more efficient recognition and classification across varied scenarios.

When Dual-Backbone and RFDAP modules are combined, FP-DETR achieves even greater improvements, with increases of 2.32%, 1.22%, and 0.39% in precision, F1 score, and mAP@50, respectively. The dual backbone provides a framework for multi-scale feature learning, while the RFDAP module enhances fine-grained details. Together, they complement each other, enriching feature expression and significantly boosting FP-DETR's ability to recognize and classify targets under diverse conditions.

4.5.2. *Dual backbone ablation experiment*

As shown in Table 6, compared with FasterNet, the serial version of Vision Mamba achieves 2.9% higher precision and 2.61% higher F1 score, with only 1.5 × 10^6 additional parameters. It also demonstrates clear advantages over ConvNeXtV2, delivering 3.78% higher precision and 3.92% higher mAP@50 with fewer parameters, as well as over EfficientViT (+5.5% precision and + 5.57% F1 score) and Swin Transformer (+3.6% precision with 11.72 × 10^6 fewer parameters).

Table 5
Key block ablation experiments.

RT-DETR	Dual-Backbone	RFDAP	Precision (%)	F1 Score	mAP@50
√			96.58	96.66	96.53
√	√		97.49	97.71	95.71
√		√	97.72	97.75	96.81
√	√	√	98.9	97.88	96.92

In parallel deployment, Vision Mamba achieves 98.9% precision, 97.88% F1 score, and 96.92% mAP@50. Compared with its serial version, precision decreases slightly by 0.1%, but F1 score improves by 0.48%, while parameters are reduced by 9.92 × 10^6, leaving only 37.3% of the serial version. This balance highlights the parallel version's suitability for lightweight and high-precision requirements in real-time edge device detection.

4.6. *Onsite experiment validation for the lightweight model of pig body temperature detection*

To evaluate the practical efficiency of FP-DETR in real-world applications, we deployed it on an inspection robot within the pig house. The robot integrates a visible–infrared dual-mode camera, a lifting mechanism, a 360° rotating pan-tilt unit, a touchscreen interface, and an automated obstacle-avoidance mobile chassis. Its overall structure is illustrated in Fig. 16.

To compare the performance of the robot system, a handheld thermal imaging camera (Fotric Model 287-L20, Fotric, Texas, USA; temperature range: 40°C–150°C, accuracy: ±2°C) was used for visual data collection. The measurements were taken under the same conditions as the inspection robot, including position, height (1.5 m), angle (30°), distance (1 m), and emissivity (0.98), to ensure the results were directly comparable, the selection of components and the corresponding technical specifications for each part of the inspection robot are presented in Table 7.

The experimental validation was conducted on May 30, 2025, at Hongsheng Building Science Research Institute, Harbin, China (Fig. 17).

The pig house is divided into four separate pens, each with a different number of pigs, totaling 17 pigs. The data collection period was from 1:30 pm (after feeding) to 4:30 pm (before feeding).

The pig body temperature measurements obtained by the FP-DETR–deployed inspection robot were compared with manual measurements taken using an infrared thermal camera (Fotric Model 287-L20, Fotric, Texas, USA; temperature range: 40°C–150°C, accuracy: ±2°C). Processing times for 17 pigs were recorded using the system's built-in clock. The automated method required 56.5 s per pig (0.94 min), whereas manual measurements took 98.8 s per pig (1.65 min), representing a 42.9% reduction in time per animal. Extrapolated to a 100-pig farm, this improvement translates into a time saving of 70.5 min per full inspection cycle, reducing the total inspection duration from 2.75 h to 1.57 h, while maintaining accuracy.

As shown in Fig. 18, the Bland–Altman analysis indicates that temperature differences are tightly clustered around the mean bias of 0.16°C, with 95% limits of agreement (LoA) ranging from –0.44°C to 0.76°C (span: 1.20°C), demonstrating an acceptable level of agreement in practice. The difference data points are distributed symmetrically on both sides of the mean line, without any evident trend of systematic deviation. Within these limits, approximately 93% of the data points (16/17) fall inside the interval, with only a single point slightly above the upper limit, further confirming the high reliability of the measurements. Notably, the distribution of differences shows no association with temperature magnitude, indicating that the agreement between the two methods remains consistent across the full measurement range.

In order to check the robustness of the FP-DETR model in real farm conditions, real-time deployment tests were carried out. The results show that the model keeps stable accuracy, with variations mostly within 2–5%, which means an absolute error of about 0.8–1.9°C when compared with a reference body temperature of 38°C. Such errors mainly appeared in extreme situations, such as heavy rain that raised humidity and lowered ambient temperature. For efficiency, the frame rate was between 50 and 60 FPS. Lower frame rates were seen mainly at device startup and after long continuous use (more than three hours). During normal operation, the system stayed stable at 55–60 FPS, which

Table 6
Different backbone replacement ablation experiments.

Model	Parallel	Serial	Precision (%)	F1Score (%)	mAP@50(%)	Parameters/10^6
FasterNet	√		96	95.27	92.66	4.4
		√	96.6	94	93.28	10.8
ConvNeXtV2	√		95.12	92.81	93	5.7
		√	96.8	94.76	96.3	12.6
EfficientViT	√		93.4	92.31	92.45	6.1
		√	92.2	93.11	92.86	11
Swin Transformer	√		95.3	94.16	94.21	17.62
		√	97.2	96.44	96.13	36.61
Vision Mamba (ours)	√		**98.9**	**97.88**	**96.92**	**5.9**
		√	**99**	**97.4**	**97.72**	**15.82**

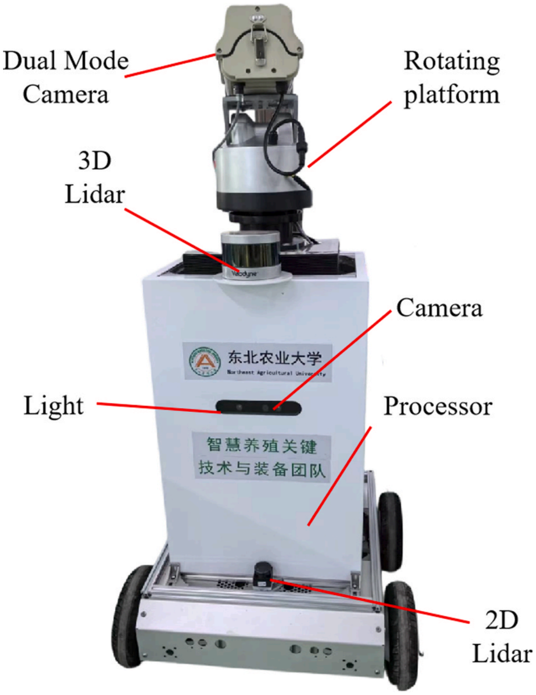

Fig. 16. The appearance of the inspection robot.

Table 7
Hardware parameters of the experimental platform.

Component	Version	Manufacturer	Parameter
3D Lidar	VLP-16	Velodyne, San Jose, CA, USA	16-line,100 m detection range;300,000 points per second
2D Lidar	MS200	ORADAR, Shenzhen, China	Single line; 12 m measurement range;4500 points per second
Dual-mode camera	UD36833B	Hikvision, Hangzhou, China	256 x 128; 28 FPS
Depth Camera	Astras	Orbbec, Shenzhen, China	640 x 480: 30 FPS
IMU	CMP10A	Yahboom, Shenzhen, China	Output frequency:0.2–200 Hz; 10 axis
Processor	Jetson-Nano	NVIDIA, Santa Clara,CA,USA	4 Core A57: 472 GFLOPs
Motion chassis	TR500	HelloMaker, Shenzhen, China	Crawler-type;0–1.2 m/s running speed

is suitable for reliable real-time monitoring.

In summary, the testing device shows high-precision temperature extraction capabilities with robust measurement agreement, indicating strong practical ability.

4.7. *Current limitations and future development*

Although FP-DETR shows excellent detection performance on the current dataset, several potential aspects for future research and improvement still exist. First, due to objective limitations, the pig breeds available in experimental farms are limited. Consequently, model validation is primarily concentrated on data from a single breed, representing an unavoidable limitation at this stage. Future access to broader farm and pig species resources would be helpful for further examining the model's ability to generalize across different species and growth stages.

Moreover, although the model has achieved a good balance between accuracy and efficiency, its deployment on edge devices remains constrained by hardware limitations. Future efforts should focus on further optimization through lightweight network design, model compression, and hardware-software co-optimization to ensure stable real-time operation in practical farming environments. In addition, exploring the integration of the model into microcontroller units (MCUs) to develop truly lightweight detection devices with low power consumption, low cost, and high portability could better meet the practical needs of on-site farming applications.

5. Conclusion

(1) FP-DETR showed higher performance than other models, achieving higher precision (99.07 ± 0.50%) than YOLOv9 (92.65 ± 0.52%), YOLOv8 (94.78 ± 0.40%) and Faster R-CNN (89.44 ± 0.49%). It ran at higher FPS (67 ± 10 FPS) than RT-DETR (41 ± 6 FPS), while maintaining higher efficiency with smaller size (16.79 ± 0.21 MB) and faster training time (8.79 ± 0.28 h).

(2) The FP-DETR detection model meets the requirement of real-time temperature extraction. The detection frame rate is 68 FPS, which is higher than the general requirement for real-time detection (25 ~ 30 FPS).

(3) The temperature extracted by the FP-DETR model is highly consistent with the manual extraction results, which can effectively replace the manual measurement. The R^2, MAE and RMSE of the extraction temperature were 0.957, 0.108 and 0.142, respectively.

(4) The FP-DETR system achieved 43% faster detection (17 pigs in 16 min) while maintaining high accuracy (0.16°C MAE), with 93% of measurements falling within acceptable consistency limits.

CRediT authorship contribution statement

Jinghan He: Writing – original draft, Methodology, Formal analysis, Data curation. **Hong Zhou:** Writing – review & editing, Data curation. **Qiuju Xie:** Writing – review & editing, Supervision, Conceptualization. **Wenwu Wang:** Supervision, Writing – review & editing. **Xuefei Liu:** Resources, Investigation, Data curation. **Wenyang Liu:** Writing – review & editing. **Yuhuan Guo:** Writing – review & editing. **Honggui Liu:**

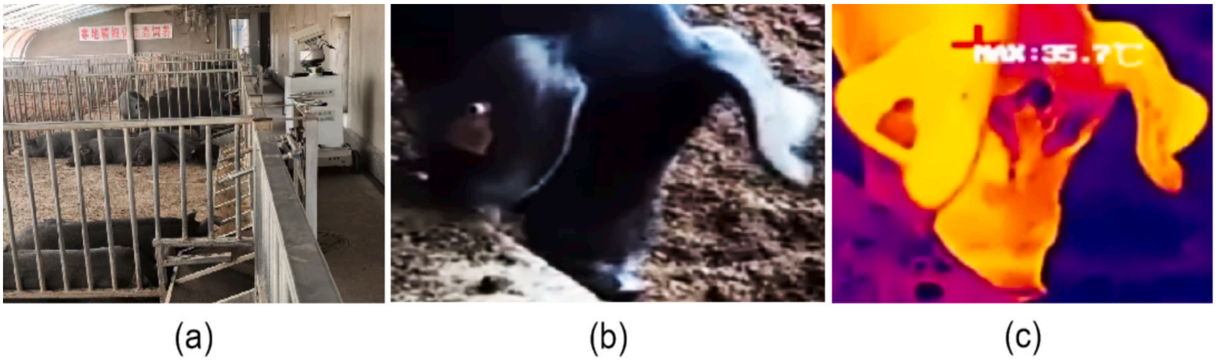

Fig. 17. Data collection equipment. (a) Data collection process. (b) Visible light image. (c) Infrared images.

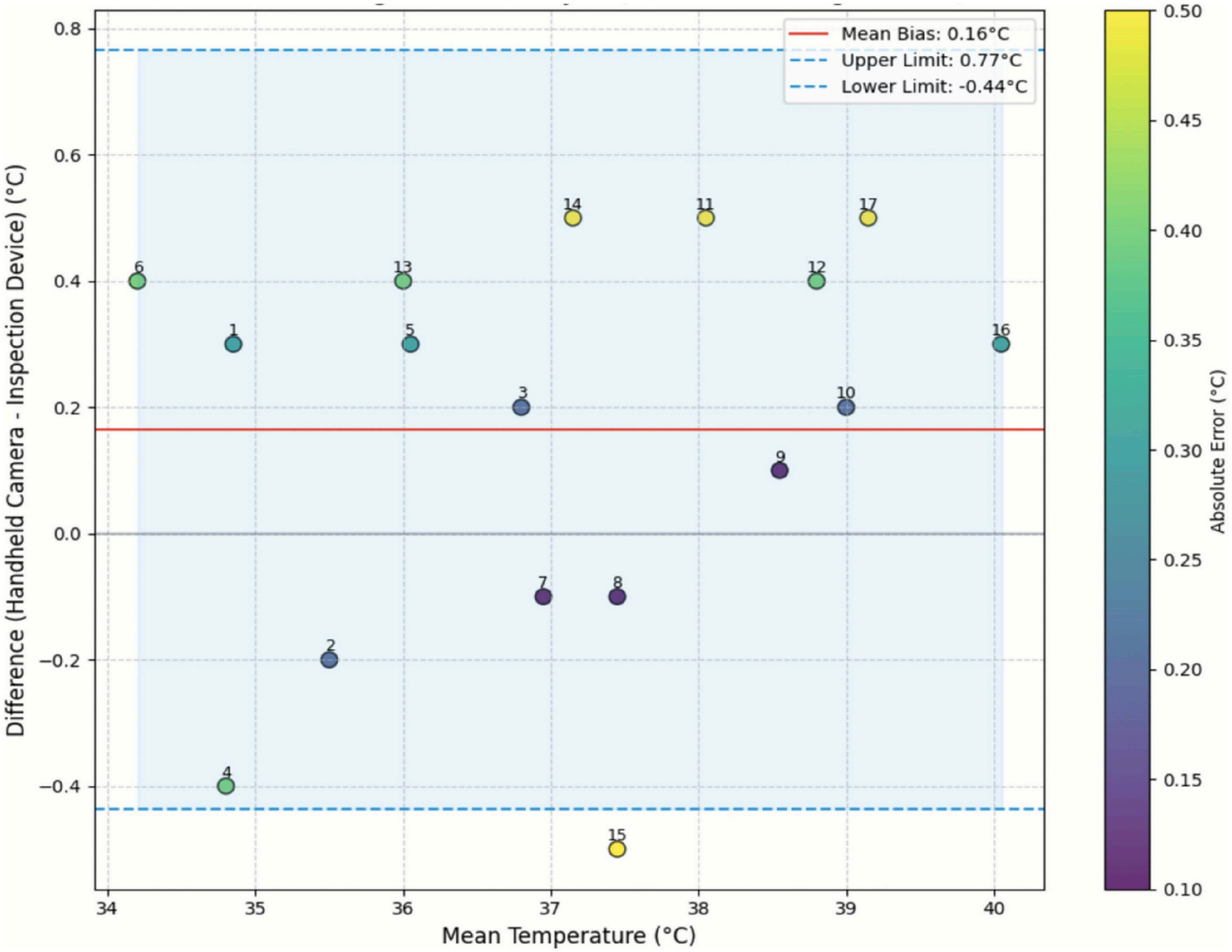

Fig. 18. Comparison of experimental results.

Supervision, Conceptualization.

Declaration of competing interest

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

Acknowledgements

This work was supported by the project of National Natural Science Foundation of China (NSFC) (32372934, 32072787); the key project of Joint Natural Science Foundation of Heilongjiang Province, China (ZL2024C017); the project of Postdoctoral Science Foundation of Heilongjiang Province, China (LBHQ21070); the project of International High Impact Paper of Northeast Agriculture University (212-54900112).

The authors acknowledge the support by the managers and workers at Jingzhe pig farm, Hongzhukangyuan pig farm, and Harbin Hongsheng Scientific Research Institute, Heilongjiang Province, China, and the help by all others for this study.

Data availability

Data will be made available on request.

References

Bagavathiappan, S., Lahiri, B.B., Saravanan, T., Philip, J., Jayakumar, T., 2013. Infrared thermography for condition monitoring–a review. Infrared Phys. Technol. 60, 35–55. https://doi.org/10.1016/j.infrared.2013.03.006.

Benjamin, M., Yik, S., 2019. Precision livestock farming in swine welfare: a review for swine practitioners. Animals 9, 133. https://doi.org/10.3390/ani9040133.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.. Encoder-decoder with atrous separable convolution for semantic image segmentation. https://doi.org/10.48550/arXiv.1802.02611.

Chen, L., Gu, L., Li, L., Yan, C. and Fu, Y., 2025, Frequency Dynamic Convolution for Dense Image Prediction, arXiv preprint arXiv:2503.18783, https://doi.org/10.48550/arXiv.2503.18783.

Chollet, F., 2017, Xception: Deep learning with depthwise separable convolutions, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258, doi: 10.1109/CVPR.2017.195.

Chung, H., Li, J., Kim, Y., Van Os, J.M., Brounts, S.H., Choi, C.Y., 2020. Using implantable biosensors and wearable scanners to monitor dairy cattle's core body temperature in real-time. Comput. Electron. Agric. 174, 105453. https://doi.org/10.1016/j.compag.2020.105453.

Cuthbertson, H., Tarr, G., González, L.A., 2019. Methodology for data processing and analysis techniques of infrared video thermography used to measure cattle temperature in real time. Comput. Electron. Agric. 167, 105019. https://doi.org/10.1016/j.compag.2019.105019.

Galina, M., Safitri, C., Bukhori, I., Silitonga, A., Suhartomo, A., 2022. An an implementation of smart agriculture for optimizing growth using Sonic Bloom and IoT integrated. J. Infotel 14, 65–74. https://doi.org/10.20895/infotel.v14i1.725.

Guo, Q., Sun, Y., Orsini, C., Bolhuis, J.E., de Vlieg, J., Bijma, P., de With, P.H., 2023. Enhanced camera-based individual pig detection and tracking for smart pig farms. Comput. Electron. Agric. 211, 108009. https://doi.org/10.1016/j.compag.2023.108009.

Hossain, S., Chowdhury, M.P.H.B., 2024. AgroSense: an IoT-based manual crops selection farming. Int. J. Information and Commun. Technol. (IJoICT) 10, 53–61. https://doi.org/10.21108/ijoict.v10i1.918.

Jumi, J., 2024. Design and building of a breeding house for IoT-based goat farming. J. Infotel. https://doi.org/10.20895/infotel.v16i3.1223.

Listianingsih, W., Susanto, T., 2023. Toward smart environment and forest city success: embracing sustainable urban solutions. Indonesian J. Computing (Indo-JC) 8, 23–34. https://doi.org/10.34818/INDOJC.2023.8.2.727.

Lohse, L., Uttenthal, Å., Enøe, C., Nielsen, J., 2010. A study on the applicability of implantable microchip transponders for body temperature measurements in pigs. Acta Vet. Scand. 52, 1–9. https://doi.org/10.1186/1751-0147-52-29.

Lu, M., He, J., Chen, C., Okinda, C., Shen, M., Liu, L., Yao, W., Norton, T., Berckmans, D., 2018. An automatic ear base temperature extraction method for top view piglet thermal image. Comput. Electronics in Agric. 155, 339–347. https://doi.org/10.1016/j.compag.2018.10.030.

Lu, Z., Zhao, M., Luo, J., Wang, G., Wang, D., 2021. Automatic teat detection for rotary milking system based on deep learning algorithms. Comput. Electron. Agric. 189, 106391. https://doi.org/10.1016/j.compag.2021.106391.

Marquez, H.P., Ambrose, D., Schaefer, A., Cook, N., Bench, C., 2019. Infrared thermography and behavioral biometrics associated with estrus indicators and ovulation in estrus-synchronized dairy cows housed in tiestalls. J. Dairy Sci. 102, 4427–4440. https://doi.org/10.3168/jds.2018-15221.

Opriessnig, T., Giménez-Lirola, L., Halbur, P., 2011. Polymicrobial respiratory disease in pigs. Anim. Health Res. Rev. 12, 133–148. https://doi.org/10.1017/S1466252311000120.

Pratama, F.D., Mutiara, G.A., Meisaroh, L., 2023. A virtual cage for monitoring system semi-intensive livestockâ€™ s using wireless sensor network and Haversine method. J. Infotel 15, 201–208. https://doi.org/10.20895/infotel.v15i2.944.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., Howard, A., 2024. MobileNetV4: Universal models for the mobile ecosystem. European Conference on Computer Vision 78-96. https://doi.org/10.48550/arXiv.2404.10518.

Ramirez, A., Karriker, L.A., 2019. Herd evaluation. Diseases of Swine 1–16. https://doi.org/10.1002/9781119350927.ch1.

Salguero, F.J., 2020. Comparative pathology and pathogenesis of African swine fever infection in swine. Front. Vet. Sci. 7, 282. https://doi.org/10.3389/fvets.2020.00282.

Sasaki, Y., Furusho, K., Ushijima, R., Tokunaga, T., Uemura, R., Sueyoshi, M., 2016. Body surface temperature of suckling piglets measured by infrared thermography and its association with body weight change. Japan Agric. Res. Quarterly: JARQ 50, 361–368. https://doi.org/10.6090/jarq.50.361.

Sellier, N., Guettier, E., Staub, C., 2014. A review of methods to measure animal body temperature in precision farming. American J. Agric. Sci. Technol. 2, 74–99. https://hal.science/hal-01512238.

Shofura, S., Suryani, S., Salma, L., Harini, S., 2021. The effect of number of factors and data on monthly weather classification performance using artificial neural networks. Int. J. Info. Commun. Technol. (IJoICT) 7, 23–35. https://doi.org/10.21108/ijoict.v7i2.602.

Tzanidakis, C., Simitzis, P., Arvanitis, K., Panagakis, P., 2021. An overview of the current trends in precision pig farming technologies. Livest. Sci. 249, 104530. https://doi.org/10.1016/j.livsci.2021.104530.

Whittaker, A.L., Muns, R., Wang, D., Martínez-Burnes, J., Hernández-Ávalos, I., Casas-Alvarado, A., Domínguez-Oliva, A., Mota-Rojas, D., 2023. Assessment of pain and inflammation in domestic animals using infrared thermography: a narrative review. Animals 13, 2065. https://doi.org/10.3390/ani13132065.

Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S., 2018. Cbam: Convolutional block attention module, Proceedings of the European conference on computer vision (ECCV), pp. 3-19 , doi: 10.48550/arXiv.1807.06521.

Xiao, D., Lin, S., Liu, Q., Huang, Y., Zeng, R., Chen, L., 2021. Automatic ear temperature extraction algorithm for live pigs based on infrared thermography. Trans. Chin. Soc. Agric. Mach 52, 255–262. https://doi.org/10.6041/j.issn.1000-1298.2021.08.026.

Xie, Q., Wu, M., Bao, J., Zheng, P., Liu, W., Liu, X., Yu, H., 2023. A deep learning-based detection method for pig body temperature using infrared thermography. Comput. Electron. Agric. 213, 108200. https://doi.org/10.1016/j.compag.2023.108200.

Zhang, X., Kang, X., Feng, N., Liu, G., 2020. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. Comput. Electron. Agric. 178, 105754. https://doi.org/10.1016/j.compag.2020.105754.

Zhang, Z., Zhang, H., Liu, T., 2019. Study on body temperature detection of pig based on infrared technology: a review. Artif. Intell. Agric. 1, 14–26. https://doi.org/10.1016/j.aiia.2019.02.002.

Zhang, B., Xiao, D., Liu, J., Huang, S., Huang, Y., Lin, T., 2024. Pig eye area temperature extraction algorithm based on registered images. Comput. Electron. Agric. 217, 108549. https://doi.org/10.1016/j.compag.2023.108549.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.. Detrs beat yolos on real-time object detection. https://doi.org/10.48550/arXiv.2304.08069.

Zhou, L., Chen, D., Chen, Z., Yuan, Y., Wang, L., Sun, X., 2017. Pig ear abnormal color detection on image processing techniques. Trans. Chinese Soc. Agric. Machinery 48, 166–172. https://doi.org/10.6041/j.issn.1000-1298.2017.04.022.