Automated collection of facial temperatures in dairy cows via improved UNet

Computers and Electronics in Agriculture

Shu, Hang; Wang, Kaiwen; Guo, Leifeng; Bindelle, Jérôme; Wang, Wensheng

https://doi.org/10.1016/j.compag.2024.108614

Computers and Electronics in Agriculture 220 (2024) 108614

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Automated collection of facial temperatures in dairy cows via improved UNet

Hang Shu [a,b,*,1], Kaiwen Wang [a,c,1], Leifeng Guo [a], Jérôme Bindelle [b], Wensheng Wang [a,*]

[a] *Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100086, China*
[b] *AgroBioChem/TERRA, Precision Livestock and Nutrition Unit, Gembloux Agro-Bio Tech, University of Liège, 5030 Gembloux, Belgium*
[c] *Information Technology Group, Wageningen University and Research, 6708 PB Wageningen, the Netherlands*

ARTICLE INFO

Keywords:
Precision livestock farming
Animal welfare
Deep learning
Body surface temperature
Heat stress

ABSTRACT

In cattle, facial temperatures captured by infrared thermography provide useful information from physiological aspects for researchers and local practitioners. Traditional temperature collection requires massive manual operations on relevant software. Therefore, this paper aimed to propose a tool for automated temperature collection from cattle facial landmarks (i.e., eyes, muzzle, nostrils, ears, and horns). An improved UNet was designed by replacing the traditional convolutional layers in the decoder with Ghost modules and adding Efficient Channel Attention (ECA) modules. The improved model was trained on our open-source cattle infrared image dataset. The results show that Ghost modules reduced computational complexity and ECA modules further improved segmentation performance. The improved UNet outperformed other comparable models on the testing set, with the highest mean Intersection of Union of 80.76% and a slightly slower but still good inference speed of 32.7 frames per second. Further agreement analysis reveals small to negligible differences between the temperatures obtained automatically in the areas of eyes and ears and the ground truth. Collectively, this study demonstrates the capacity of the proposed method for automated facial temperature collection in cattle infrared images. Further modelling and correction with data collected in more complex conditions are required before it can be integrated into on-farm monitoring of animal health and welfare.

1. Introduction

Infrared thermography (IRT) is the technique of detecting infrared radiation from an object, converting it to temperature, and visualizing the temperature distribution with an image (Tan et al., 2009). Due to its non-contact advantage, IRT has been widely used in human fever detection and health evaluation. In husbandry, IRT can contribute to precision livestock farming which aims to provide an automated protocol for monitoring animal health and welfare parameters (Halachmi and Guarino, 2016).

In cattle, temperatures obtained from specific facial landmarks (e.g., eyes, forehead, nostril, ears, horns, cheek) have been widely used as indicators or predictors of health conditions such as physiological state (Ma et al., 2021), bovine respiratory disease (Schaefer et al., 2012), and foot-and-mouth disease (Gloster et al., 2011); animal welfare issues such as temperament (Chen et al., 2021), emotions (Uddin et al., 2021), and heat stress (Peng et al., 2019); and productivity issues such as feed efficiency (Montanholi et al., 2009) and meat quality (Cuthbertson et al., 2020).

In order to collect the temperature of the abovementioned facial landmarks, facial regions of interest (ROIs) must first be defined. In most literature, ROIs are defined manually in infrared images using relevant processing software due to the lack of reliable detection tools for cattle facial landmarks (Cuthbertson et al., 2019; Lowe et al., 2019). Thus, there has been a growing interest among researchers to develop such a tool to increase the efficiency of dealing with cattle infrared images. In previous studies, ROIs such as eyes, ear base, cheek, and nose have been localised in infrared images for specific purposes (Jorquera-Chavez et al., 2019; Kim and Hidaka, 2021; Lowe et al., 2020; Zhang et al., 2020). However, very limited effort has been contributed yet to a comprehensive method for separating multi-class facial ROIs in cattle infrared images. This method should be robust against usual interfering factors such as camera angle and changing microenvironment.

Of note, most of the previous works use traditional image processing

* Corresponding authors at: Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100086, China (H. Shu).
E-mail addresses: shuhang@caas.cn (H. Shu), wangwensheng@caas.cn (W. Wang).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.compag.2024.108614
Received 24 June 2023; Received in revised form 12 November 2023; Accepted 2 January 2024
Available online 10 April 2024

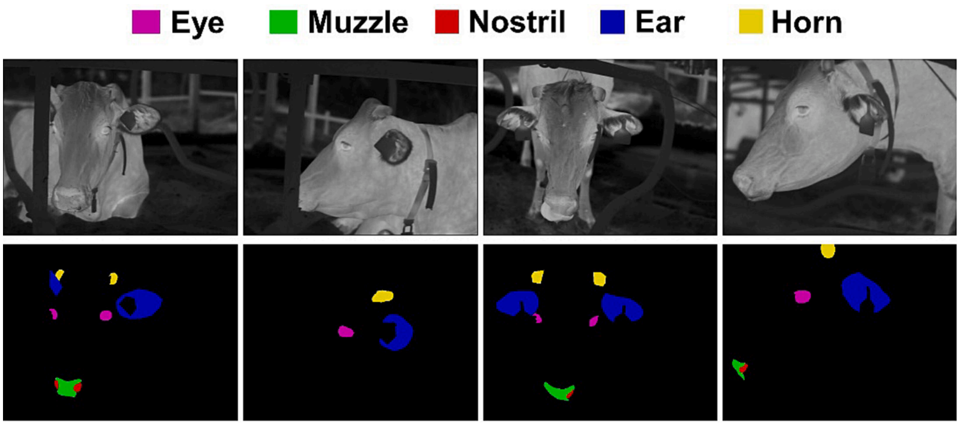

Fig. 1. Examples of cattle infrared images and their ground-truth annotations for facial landmarks.

techniques for detecting facial landmarks in cattle infrared images, such as Haar cascade classifiers (Lowe et al., 2020) and thresholding (Jaddoa et al., 2021). The recent development of deep learning provides alternative solutions. For example, recent studies have achieved automatic ocular temperature collection using object detection methods based on improved YOLOv4 (Wang et al., 2022a) and YOLOv5 (Wang et al., 2022b), and YOLOv7 (Chu et al., 2023). In addition, semantic segmentation, as another central computer vision task that separates each pixel into pre-defined classes (LeCun et al., 2015). Its usefulness is particularly evident in the field of face parsing, which aims to create pixel-wise segmentation maps for facial parts in human RGB images (Lin et al., 2019). It can be therefore imagined that this pixel-wise outlining can lead to more accurate and comprehensive temperature assessments from infrared images. However, no relevant work or attempts have been found yet.

Semantic segmentation in infrared images has to deal with some challenges (Kütük and Algan, 2022). One of the main challenges is the lower resolution of infrared images compared with RGB images, which can result in less detailed information and difficulty in accurately delineating object boundaries. Another challenge is the phenomenon of thermal crossover, where objects at similar temperatures blend into the background, making it harder to distinguish them. Advancements of more sophisticated algorithms and deep learning models, such as convolutional neural networks (CNNs), Vision Transformer, and other attention mechanisms, have markedly improved accuracy in recognition across various data, such as speech (Khan et al., 2023a) and image (Khan et al., 2023b). Therefore, it is of great interest to explore to what extent these techniques can help in segmenting facial landmarks in cattle infrared images.

Thus, this study aimed to propose a semantic segmentation-based tool for automated facial temperature collection from cattle infrared images, so as to improve the efficiency of processing relevant research data. Specifically, a baseline semantic segmentation network, namely UNet, was modified, trained, and compared its performance with other state-of-the-art models in segmenting cattle facial landmarks. Then, the temperatures obtained from the predictions of the improved UNet were compared with those obtained from the ground-truth annotations.

2. Materials and methods

Since there is no public infrared image dataset appropriate for the semantic segmentation of cattle facial landmarks, a field experiment was conducted for data acquisition. The experimental protocol was approved by the Experimental Animal Care and Committee of Institute of Animal Sciences, Chinese Academy of Agricultural Sciences (approval number IAS2021-220).

2.1. *Data acquisition*

The experimental farm is located in Shandong, China (34∘50′37″N and 115∘26′11″E), and belongs to a temperate continental monsoon climate with hot and humid summers. It is worth noting that the temperature difference between the background and animals would change dramatically from non-heat-stressed months to heat-stressed months. Ignoring this fact would definitely affect the robustness of the trained network in practice. Thus, the experiment was conducted from May to August 2021 to cover a wide range of thermal environments from warm to hot. The free-stall pen was covered by a double-pitched roof, and therefore, most of the solar radiation was prevented from reaching the cows inside the barn. Indoor microenvironmental parameters including ambient temperature (Ta) and relative humidity (RH) were measured by using six Kestrel 5000 and 5400 environment meters that were equally spaced in the barn (measurement interval: 10 min, accuracy: ± 0.4 °C Ta and ± 1 % RH; Nielsen-Kellerman, Boothwyn, PA, USA). The temperature-humidity index (THI) was calculated according to Eq. (1) (NRC, 1971).

$$THI = (1.8 \times Ta + 32) - (0.55 - 0.005 \times RH) \times (1.8 \times Ta - 26) \tag{1}$$

A total of 59 primiparous and multiparous Holstein dairy cows were selected for infrared thermal imaging. The infrared images were taken with a portable infrared camera (VarioCAM HR, InfraTec, Dresden, Germany) which has a spectral range from 7.5 to 14 μm, a temperature measuring range from − 40 °C to 2,000 °C, an accuracy of ±1.5 °C, and a resolution of 640 × 480 pixels. All images were taken at a distance of approximately 1 to 1.5 m from the cow. To increase the robustness of the proposed method in actual farms, cows were not restrained during photography and a wide range of situations including heterogeneous postures were covered. Thermal imaging was carried out between 08:00 and 17:00 h. All cows were healthy during the entire experiment.

2.2. *Data pre-processing*

Infrared images were initiated with IRBIS 3 Standard software (YSHY, Beijing, China). Before formal processing, images with low quality and multiple faces were manually eliminated, contributing to a dataset with 1,000 images. All images were calibrated by setting the emissivity to 0.98 (Montanholi et al., 2015), and by inputting the averaged Ta record from the sensors corresponding to the time when they were taken. The images were outputted into grey-scale joint photographic experts group (JPEG) format (640 × 480 pixels) with the temperature scale set to 295 to 315 degrees Kelvin. The temperature matrices were also outputted into comma-separated values (CSV) format for further temperature collection from segmentations.

For a given grey-scale image, facial landmarks that have been

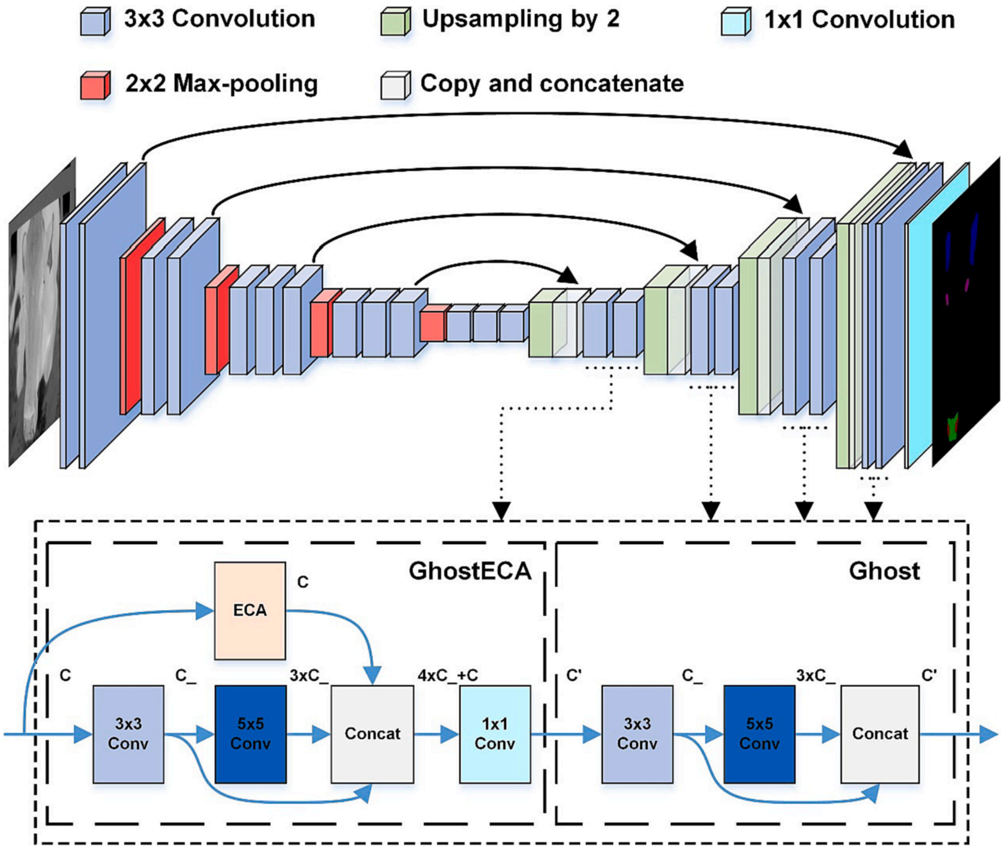

Fig. 2. Architecture of the baseline UNet model and the improved UNet model by replacing convolutions in the decoder with Ghost and GhostECA modules (shown in dashed boxes, where C and C' represent the input and output channel numbers, respectively, and C_=C'/4).

frequently used in dairy research (i.e., eyes, muzzle, nostrils, ears, horns) were annotated with polygons using Labelme (https://github.com/wkentaro/labelme.git) (Fig. 1). The images were randomly allocated to training (46 cows, 782 images), validation (6 cows, 125 images), and testing (7 cows, 93 images) sets. The training set was used to train the networks, the validation set to tune the hyperparameters and obtain an initial assessment of accuracy, and the testing set to collect the final performance. Data augmentation methods, such as flipping, rotation, brightness changing, contrast changing, sharpening, Gaussian noise adding, and elastic deformation, were performed in the training set to improve the accuracy and generalisation capacity of the trained network (see Appendix A. Supplementary Material Fig. S1). Thus, the training set was seven-fold augmented. The images were resized to 512 × 512 pixels before they were fed to the networks.

2.3. *Segmentation network architectures*

2.3.1. *UNet model*

UNet, as a popular semantic segmentation network, has a symmetric U-shaped architecture of a contracting path for capturing global context and an expansive path for precise localisation, and uses skip connections between two paths to transfer context information to higher resolution layers (Ronneberger et al., 2015). In this study, the UNet with a VGG16 (Simonyan and Zisserman, 2014) encoder was used as the baseline (Fig. 2). The downsampling block is repeated by two or three 3 × 3 convolutions (activated by ReLU functions) and one 2 × 2 max-pooling operation. Thus, the image is halved in size after each block, compensated with a doubled number of feature map channels. In the decoder part, repeated blocks include an upsampling (bilinear method), a concatenation with the corresponding feature map from the encoder, and two repeated 3 × 3 convolutions (each followed by a ReLU function) to fuse and reconstruct feature maps from both local details and global context. Finally, a 1 × 1 convolution with the number of channels set to the number of classes is used to generate class-wise classification results for each pixel.

2.3.2. *Improved UNet model*

As shown in Fig. 2, the modification on the baseline UNet model happened to the decoder where two consecutive convolutional layers are replaced by a combination of a Ghost with Efficient Channel Attention (GhostECA) module and a Ghost module. The detailed structure of the improved UNet is shown in Table 1.

The Ghost module is a plug-and-play component that can be used to replace common convolutional operations in any classical CNNs (Han et al., 2020). The idea of Ghost modules was from the observation of the intermediate feature maps calculated by mainstream CNNs. The authors found the redundancy in feature maps, in which some feature maps are very similar in pairs, as if one of the pair is a “ghost” of the other. This means that one feature map of the pair can be obtained by transforming the other feature map with cheap operations. Therefore, Ghost modules aim to generate more feature maps with fewer parameters and cheaper operations.

The applied Ghost module, as shown in Fig. 2, consists of two consecutive convolutions. In the first convolution, the number of input channels is reduced to one-quarter, using a kernel size of 3 and a stride of 1. This reduction in channel dimensions helps reduce model complexity and computational cost. The second convolution expands the reduced channels by three times, using a larger kernel size of 5 and a stride of 1. Finally, the output of the second convolutional operation is concatenated with that of the first convolution to assure that the number of output channels equals the number of original input channels.

It is well known that adding attention mechanisms to CNNs can

Table 1
Structure of the improved UNet.

Layer	Kernel size & stride	Output shape	Connect to
Input	–	512 × 512 × 3	Convolution1
Convolution1	3 × 3, 1	512 × 512 × 64	Convolution2
Convolution2	3 × 3, 1	512 × 512 × 64	Max-pooling1 & Concatenate4
Max-pooling1	2 × 2, 2	256 × 256 × 64	Convolution3
Convolution3	3 × 3, 1	256 × 256 × 128	Convolution4
Convolution4	3 × 3, 1	256 × 256 × 128	Max-pooling2 & Concatenate3
Max-pooling2	2 × 2, 2	128 × 128 × 128	Convolution5
Convolution5	3 × 3, 1	128 × 128 × 256	Convolution6
Convolution6	3 × 3, 1	128 × 128 × 256	Convolution7
Convolution7	3 × 3, 1	128 × 128 × 256	Max-pooling3 & Concatenate2
Max-pooling3	2 × 2, 2	64 × 64 × 256	Convolution7
Convolution8	3 × 3, 1	64 × 64 × 512	Convolution8
Convolution9	3 × 3, 1	64 × 64 × 512	Convolution9
Convolution10	3 × 3, 1	64 × 64 × 512	Max-pooling4 & Concatenate1
Max-pooling4	2 × 2, 2	32 × 32 × 512	Convolution9
Convolution11	3 × 3, 1	32 × 32 × 512	Convolution10
Convolution12	3 × 3, 1	32 × 32 × 512	Convolution13
Convolution13	2 × 2, 2	32 × 32 × 512	Upsampling1
Upsampling1	–	64 × 64 × 512	Concatenate1
Concatenate1	–	64 × 64 × 1024	GhostECA1
GhostECA1	–	64 × 64 × 512	Ghost1
Ghost1	–	64 × 64 × 512	Upsampling2
Upsampling2	–	128 × 128 × 512	Concatenate2
Concatenate2	–	128 × 128 × 768	GhostECA2
GhostECA2	–	128 × 128 × 256	Ghost2
Ghost2	–	128 × 128 × 256	Upsampling3
Upsampling3	–	256 × 256 × 256	Concatenate3
Concatenate3	–	256 × 256 × 384	GhostECA3
GhostECA3	–	256 × 256 × 128	Ghost3
Ghost3	–	256 × 256 × 128	Upsampling4
Upsampling4	–	512 × 512 × 128	Concatenate4
Concatenate4	–	512 × 512 × 192	GhostECA4
GhostECA4	–	512 × 512 × 64	Ghost4
Ghost4	–	512 × 512 × 64	Convolution14
Convolution14	1 × 1, 1	512 × 512 × 6	–

improve their performance. The attention mechanism in deep learning works similarly to human selective visual attention in that both aim to identify and emphasise the most important information from large amounts of data. Efficient channel attention (ECA) is an extremely efficient and lightweight channel attention mechanism proposed by Wang et al. (2020). The applied ECA module consists of three main steps (Fig. 3). Firstly, a global average pooling operation is applied to the input feature maps, squeezing the spatial dimensions W × H to 1 × 1 while retaining channel-wise information. Next, a one-dimensional convolution with a kernel size of 3 is performed to achieve local cross-channel interaction and capture channel-wise dependencies. A sigmoid activation function is then used to compute channel-wise attention weights. Finally, the attention weights are multiplied element-wise with the input feature maps, allowing the network to selectively emphasise relevant information.

In this study, an ECA module is integrated into the first Ghost module of each upsampling block in the decoder to enhance its performance (Fig. 2). The integrated GhostECA module has one more convolutional operation after concatenating the outputs of the first two convolutions and the ECA module in order to adjust the number of output channels.

2.4. Segmentation network training

The training was performed in Python 3.7 language with Pytorch 1.13.0 on a 64-bit Windows 11 computer with NVIDIA GeForce RTX 3090 GPU. Transfer learning can significantly reduce the number of required images and increase training efficiency compared with training from scratch with randomly initialised weights. In this study, the initialised weights of all encoders were transferred from the networks pre-trained on the ImageNet dataset (Deng et al., 2009). The epoch was set to 300, the batch size to 16, and the learning rate to 0.0001 with an Adam optimiser. A combination of cross-entropy and dice coefficient was used as the loss function ($L_{CE} + L_{Dice}$), as defined in Eqs. (2) and (3). Dice loss was used because it can effectively handle the pixel imbalance between foreground and background.

$$L_{CE} = -\sum_{i=1}^{C} t_i \log p_i \quad (2)$$

$$L_{Dice} = 1 - \frac{2TP}{2TP + FP + FN} \quad (3)$$

where C takes 5, indicating five classes of interest (i.e., “eye”, “muzzle”, “nostril”, “ear”, “horn”), t_i and p_i are the ground truth and the Softmax probability of each pixel for each class i, respectively, TP denotes true positive (pixels correctly classified as a class of interest), FP denotes false positive (pixels incorrectly classified as a class of interest), TN denotes true negative (pixels correctly classified as the background), and FN denotes false negative (pixels incorrectly classified as the background or a wrong class). The model from the epoch with the lowest validation loss was used for testing.

2.5. Ablation and comparison studies

Ablation tests were conducted: (1) UNet with VGG16 as the backbone was used as the baseline model; (2) based on the baseline model, the convolutional layers in the decoder were replaced by Ghost modules. This model is referred to UNet + Ghost; and (3) based on UNet + Ghost, ECA was integrated into the first Ghost module of each decoder. This is the proposed model to be compared, which is referred to UNet + GhostECA.

To show the competitiveness of the improved UNet model, it was compared with other popular semantic segmentation models in the field, including FCN (Long et al., 2015) with VGG16 as the backbone (FCN-VGG16), PSPNet (Zhao et al., 2017) with MobileNetV2 (Sandler et al., 2018) and ResNet50 (He et al., 2016) as the backbone, respectively (PSPNet-MobileNetV2 and PSPNet-ResNet50), DeepLabV3+ (Chen et al., 2018) with MobileNetV2 (Sandler et al., 2018) and Xception (Chollet, 2017) as the backbone, respectively (DeepLabV3 + -MobileNetV2 and DeepLabV3 + -Xception), UNet with ResNet50 as the backbone (UNet-ResNet50), as well as SegFormer (Xie et al., 2021) with B5 as the backbone (SegFormer-B5).

2.6. Performance evaluation

The per-class segmentation results were shown using the Intersection over Union (IoU), Recall, and Precision, as expressed in Eqs. (4–6). The IoU is the intersection of the prediction and ground truth divided by

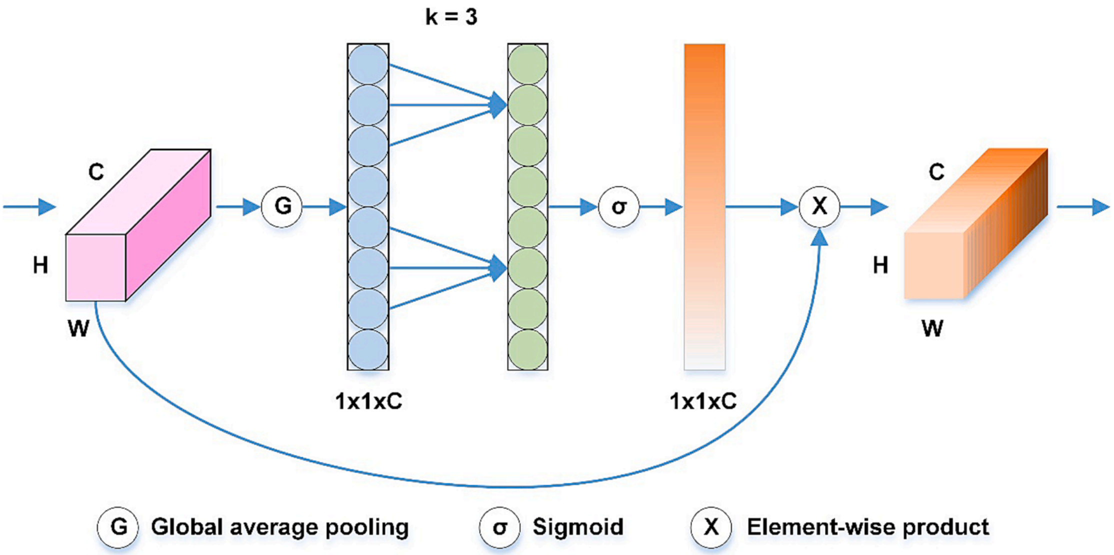

Fig. 3. Architecture of the efficient channel attention (ECA) module.

their union. The Recall indicates the proportion of all positive labels that are classified correctly. The Precision indicates the proportion of all positive predictions that are classified correctly.

$$IoU = \frac{TP}{TP + FP + FN} \times 100\% \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{6}$$

The overall segmentation performance was evaluated using the mean Intersection over Union (mIoU) and mean pixel accuracy (mPA). The mIoU is the mean IoU of the background and five classes of interest, whereas the mPA is the average of the pixel accuracy of the background and five classes, as expressed in Eqs. (7) and (8), respectively:

$$mIoU = \frac{1}{C + 1}\sum_{i=0}^{C} IoU \times 100\% \tag{7}$$

$$mPA = \frac{1}{C + 1}\sum_{i=0}^{C} \frac{p_{ii}}{\sum_{j=0}^{C} p_{ij}} \times 100\% \tag{8}$$

where C + 1 equals 6 indicating the background and five classes of interest. P_{ii} and P_{ij} are the total numbers of pixels belonging to class i that are predicted to belong to i and j, respectively.

In addition, number of parameters, model size, and floating-point operations (FLOPs) were calculated to show model complexity and computational requirements, and frames per second (FPS) was calculated to indicate the inference speed, as expressed in Eq. (9):

$$FPS = \frac{N}{t_N} \tag{9}$$

where t_N is the total inference time (s) on N images.

2.7. Data analysis

Further data analysis was done using the images from the testing set. The predicted segmentations by the improved UNet as well as the ground-truth annotations were used for generating temperature parameters (i.e., mean and maximum) of the ROIs. This was done by using a self-written program in Python that maps the coordinate matrices of the segmentations and annotations with the original temperature matrices. The results of the ground-truth annotations were regarded as ground truth. Moreover, the temperature parameters were also generated by the traditional method with ROIs manually defined on the software using appropriate shapes such as ellipses and rectangles (see Fig. S2). This was to represent the common practice in relevant studies. Finally, the results obtained by the two methods, namely the proposed method based on automated segmentations and the traditional method based on manual collection, were examined for their agreement with the ground truth using Bland-Altman plots (Altman and Bland, 1983).

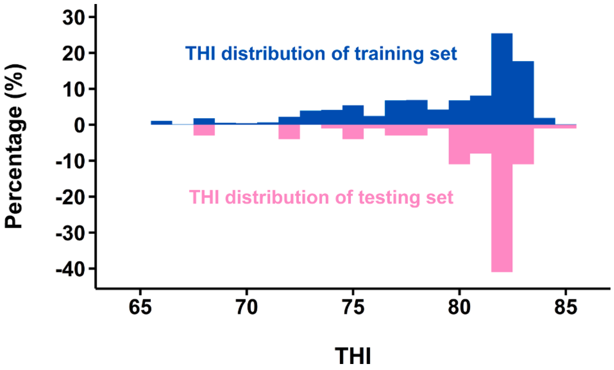

Fig. 4. Distribution of temperature-humidity index (THI) during photography summarised by training (including validation and before augmentation, n = 907) and testing (n = 93) sets, respectively.

3. Results and discussion

3.1. Overview of the dataset

During the experimental period, Ta averaged 30.1 °C (range from 22.4 to 37.6 °C), RH averaged 61.1 % (range from 19.5 to 94 %), and THI averaged 79.8 (range from 70.3 to 85.9). The standard deviation of daily mean Ta, RH, and THI were 2.7 °C, 16.3 %, and 3.2, respectively. The THI distribution shown in Fig. 4 indicates a good consistency between training and testing sets as well as wide coverage of thermal environments. According to the THI threshold customised for high-producing dairy cows, heat stress occurs at a THI of 68, mild-moderate at 72, moderate-severe at 80, and severe at 90 (Collier et al.,

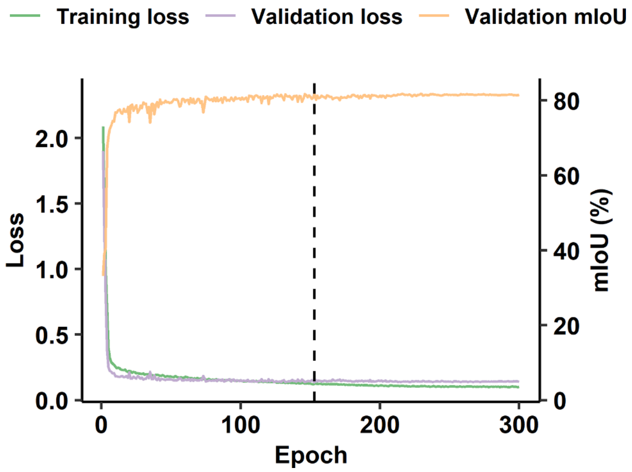

Fig. 5. Loss and mean Intersection over Union (mIoU) curve of the proposed model. The dashed line shows the epoch (153) with the lowest validation loss (0.137123).

2012). Thus, our test cows experienced no to moderate-severe heat stress during observations.

3.2. Results of training and ablation study

To the best of our knowledge, this is the first attempt at pixel-level facial landmark segmentation in cattle infrared images. As shown in Fig. 5, there was a rapid increase in validation mIoU as loss decreased at the early stage of training, and the model converged at the middle stage. The lowest loss (0.137123) on the validation set was obtained at epoch 153.

The ablation study showed a performance gain by introducing Ghost and GhostECA modules (Table 2). By replacing convolutions in the decoder with Ghost modules, the UNet + Ghost model had an increased mIoU and mPA by 0.91 % and 0.46 %, respectively, compared with the baseline UNet model. This can be explained by the enlarged receptive field as a result of the addition of larger convolutional kernels in the decoder. Plus, its number of parameters, model size, and FLOPs decreased by 30.49 %, 31.25 %, and 47.49 %, respectively, resulting in a slight increase in FPS by 3.89 %. These results are consistent with previous reports that Ghost modules can reduce computational complexity by exploiting redundancy in intermediate feature maps calculated by mainstream CNNs (Wang et al., 2022a; Zheng et al., 2023).

By further integrating the ECA module into Ghost modules, the UNet + GhostECA model further increased mIoU by 0.68 % at the basis of the UNet + Ghost model, compensated by an increased number of parameters, model size, and FLOPs by 6.53 %, 6.52 %, and 13.65 %, respectively. Still, the proposed UNet + GhostECA model outperformed the baseline model in all metrics except for FPS, which decreased by 9.17 %. These results are as expected, as the integration of attention mechanisms increases performance by suppressing the gradient transmission of irrelevant information, but often requires more computational resources due to the more complex structure. Collectively, the UNet + GhostECA model should be considered a successful improvement due to its leading mIoU, smaller computational requirements, and good inference speed.

3.3. Segmentation results of the improved UNet

The detailed performance of the improved UNet model illustrated in Fig. 6 shows robust segmentations against usual interfering factors including camera angle and extreme ambient environment. More importantly, all ROIs yielded an IoU higher than 50 % which is a commonly used threshold above which a result is considered to be accurate. The IoU, Recall, and Precision shared a similar trend, with the best performance obtained by “ear” and “eye”, while the worst by “muzzle”, “horn”, and “nostril”. The relatively poor segmentation in the nose areas was most likely due to the misclassification of pixels between nostrils and muzzles. Since the nose area of cattle is often covered by foreign matter such as mud, water, and saliva, especially during hot seasons (Burhans et al., 2022), the detection of nostrils and muzzle is more difficult than other ROIs. However, a better segmentation can be speculated by combining “muzzle” and “nostril” as one unified label class of nose areas.

The misclassification of eyes was primarily due to partially open or closed eyes. Unfortunately, this is hard to solve due to relatively limited negative samples in the current datasets. The worse results of horns can be explained by relatively fewer training instances since not all cows had horns. Also, their misclassification is partially due to incomplete horn removal. Calves were disbud using hot iron on the experimental dairy farm at around 40 days of age. If the horn bud remained subdermal, such skin surface would show a blur region on the infrared image and could be misclassified as a horn, especially from certain side views. This finding suggests a novel method for veterinarians to confirm the effectiveness of horn removal and determine whether a second operation is required.

3.4. Comparison with other semantic segmentation models

The result of the comparison study shown in Table 3 demonstrates the highest mIoU (80.76 %) by the proposed UNet + GhostECA model. The trade-off between segmentation performance and inference speed is obvious. Overall, more complex and deep networks (such as UNet and DeepLabV3 +) and backbones (such as Xception and ResNet50) had better performance metrics but lower inference speed compared with lightweight networks (such as PSPNet) and backbones (such as MobileNetV2). The only exception was SegFormer-B5 which had the largest model size and complexity but performed almost the worst results. It should be noted that SegFormer, as a transformer-based framework, was pre-trained on a dataset with cityscapes as classes of interest. On the contrary, UNet was primarily designed for segmenting medical images which are similar to our grey-scale images. This may explain why SegFormer performed worse than UNet on our infrared dataset. Moreover, DeeplabV3+, as a recent network, typically has better segmentation results when dealing with challenging tasks but performed poorly than UNet on our dataset. The poor performance of more recent and complex models (such as DeeplabV3 + and SegFormer) could be attributed to our relatively few and simple images. Indeed, UNet architecture has been reported to be more appropriate for training with limited training images and fewer deep-level features (Zou et al., 2021).

The segmentation examples shown in Fig. 7 confirm the better performance of the proposed UNet + GhostECA model. It can be seen that all models have good segmentation ability when the Ta was much lower than cattle body surface temperatures and the contour of the ROIs was obvious. However, when the Ta increased closely to cattle body surface

Table 2
Performance of ablation study on the testing set (n = 93).

Model	Backbone	mIoU (%)	mPA (%)	Number of parameters (M)	Model size (MB)	FLOPs (G)	FPS
UNet	VGG16	79.17	87.82	24.89	96	451.81	36
UNet + Ghost	VGG16	80.08	88.28	17.3	66	237.24	37.4
UNet + GhostECA (proposed)	VGG16	80.76	88.92	18.43	70.3	269.63	32.7

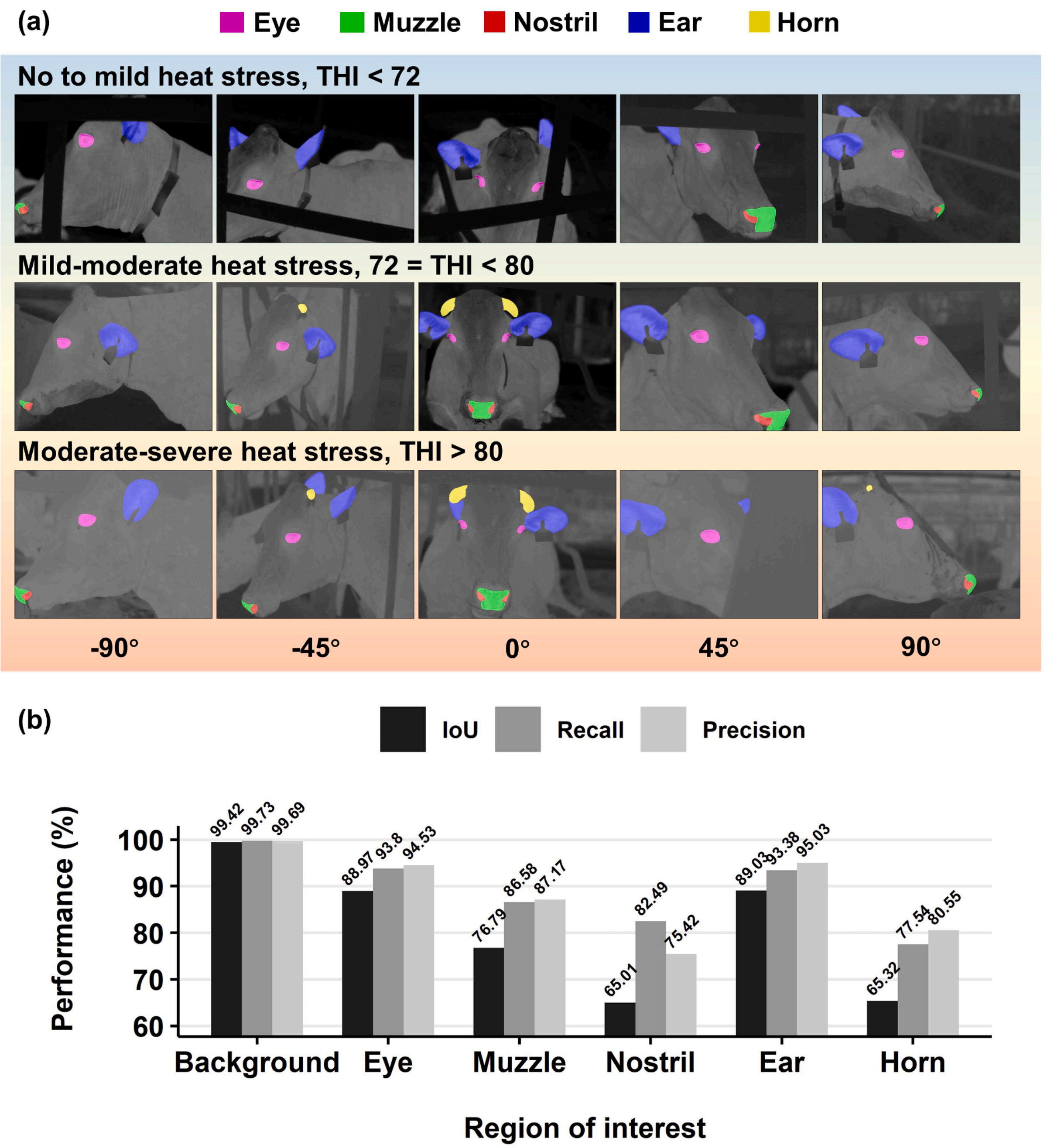

Fig. 6. Detailed segmentation results on the testing set (n = 93). (a) Predictions on some example images against varying thermal conditions classified by temperature-humidity index (THI), shown at camera angles ranging from −90° to 90°. (b) Per-class Intersection over Union (IoU), Recall, and Precision.

Table 3
Performance of different semantic segmentation models on the testing set (n = 93).

Model	Backbone	mIoU (%)	mPA (%)	Number of parameters (M)	Model size (MB)	FLOPs (G)	FPS
FCN	VGG16	76.64	84.38	19.17	73.1	204.34	45.2
PSPNet	MobileNetV2	73.73	83.38	2.38	9.3	6.03	141.4
PSPNet	ResNet50	78.82	87.4	46.71	178	118.43	85.1
DeepLabV3+	MobileNetV2	77.84	88.76	5.82	22.4	52.9	88.1
DeepLabV3+	Xception	79.14	90.33	54.71	209	166.88	28.8
UNet	VGG16	79.17	87.82	24.89	96	451.81	36
UNet	ResNet50	78.85	90.94	43.93	167	184.23	46.8
SegFormer	B5	70.93	81.49	84.6	969	986.48	21.8
Proposed	VGG16	80.76	88.92	18.43	70.3	269.63	32.7

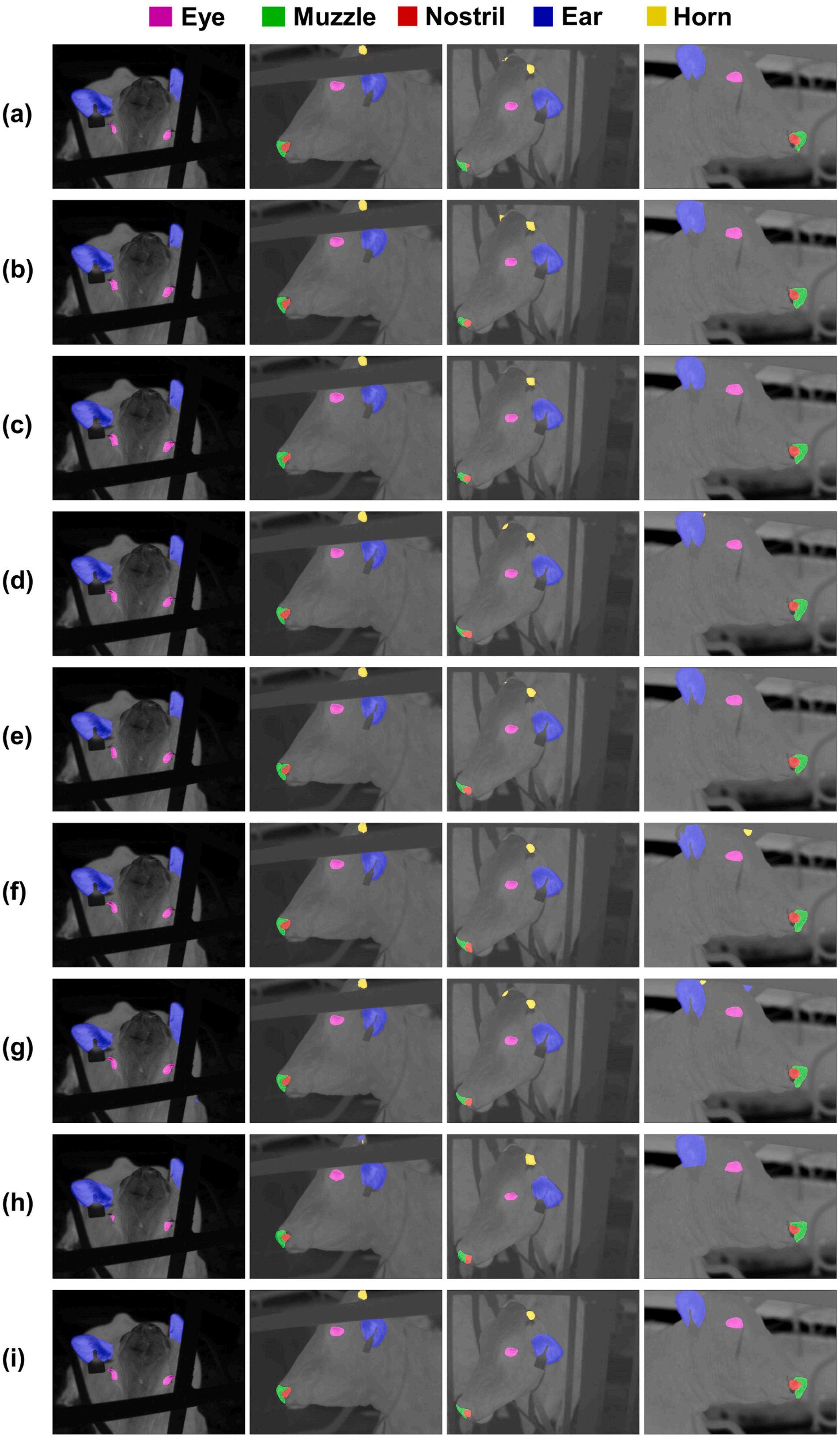

Fig. 7. Segmentation results of different semantic segmentation models on the testing set (n = 93). (a) FCN-VGG16; (b) PSPNet-MobileNetV2; (c) PSPNet-ResNet50; (d) DeepLabV3 + -MobileNetV2; (e) DeepLabV3 + -Xception; (f) UNet-VGG16; (g) UNet-ResNet50; (h) SegFormer-B5; (i) Proposed.

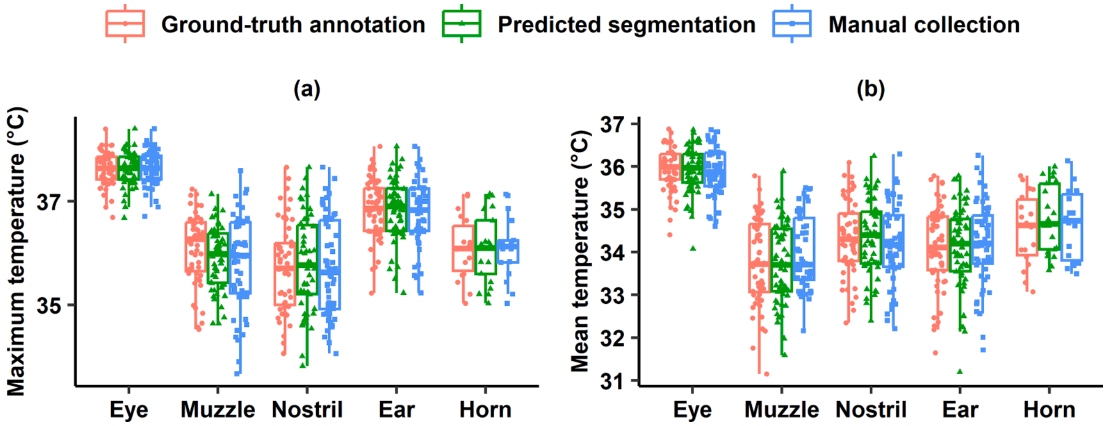

Fig. 8. Overview of the (a) maximum and (b) mean temperature of ground-truth annotation, predicted segmentation, and manual collection on the testing set (n = 93).

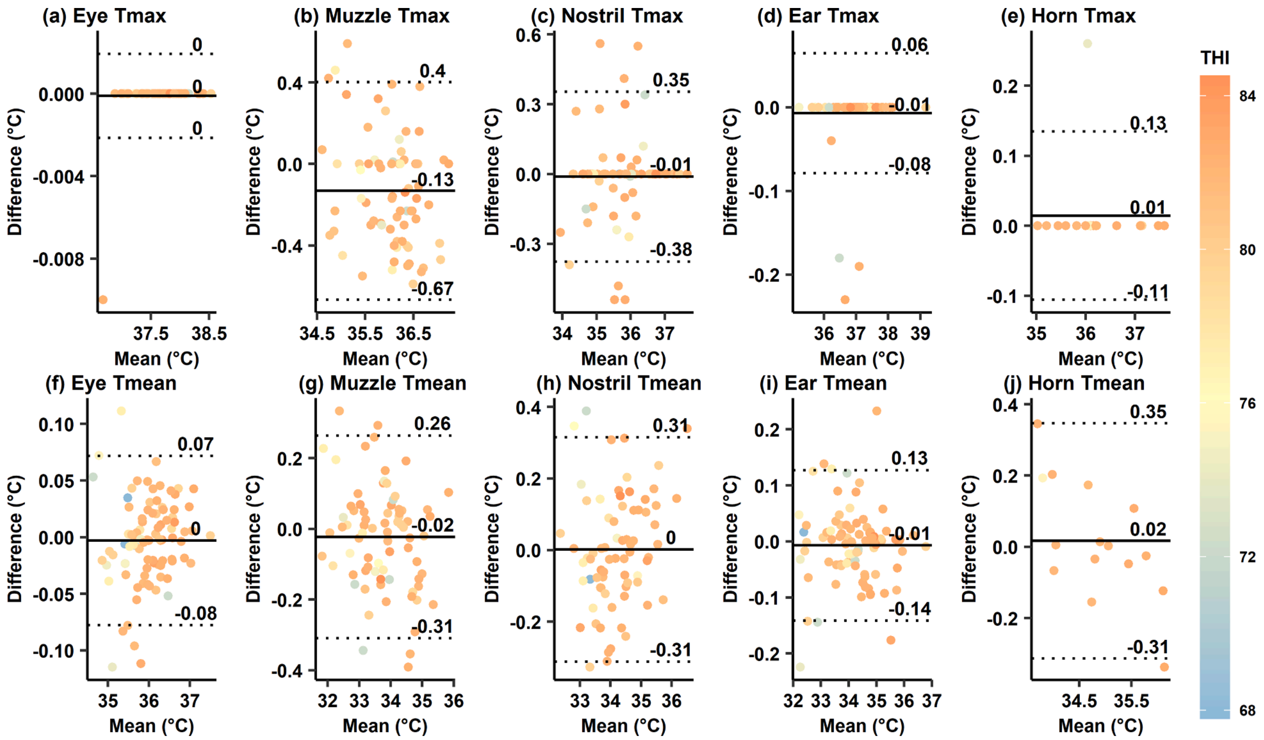

Fig. 9. Bland-Altman plots showing the agreement between predicted segmentation and ground-truth annotation on the testing set (n = 93) in terms of the maximum and mean temperatures (Tmax and Tmean) of five facial landmarks. The solid and dashed lines represent mean difference and 95 % limits of agreement, respectively. Datapoints are coloured by temperature-humidity index (THI).

temperatures and the boundary between cattle and their environments became blurry, some comparable models became less effective in segmenting the pixels at the edge while the UNet + GhostECA model still had smooth and precise segmentation. For example, the UNet + GhostECA model was the only model to precisely segment ears from ear tags in all cases, even during high Ta conditions.

3.5. Agreement of automated and manual methods with ground truth

The UNet + GhostECA model was used for further automated temperature collection due to its outperforming segmentation performance. The temperature results of the proposed automated method and traditional manual method both shared similar distributions with the ground truth, where eye temperature always had the highest values (Fig. 8). This is consistent with previous knowledge that eye temperature is the closest proxy of core body temperature among other candidate body surface temperatures (Gloster et al., 2011; Hoffmann et al., 2013).

As shown in Fig. 9, the mean differences between the temperatures obtained automatically and the ground truth are small, particularly in eyes and ears. More importantly, the differences between the temperatures obtained automatically and the ground truth (Fig. 9) had narrower limits of agreement in most cases compared with those between the temperatures obtained manually and the ground truth (Fig. 10), indicating a generally better agreement. This is reasonable since manual collection using professional software can only achieve rough coverage of the ROIs rather than pixel-wise segmentation. This common practice of collecting temperatures manually may work for maximum temperatures due to being less impacted by the overall pixels, as well as for landmarks with fewer obstacles and typical outlines (e.g., eyes). However, it can suffer when collecting mean temperatures at other

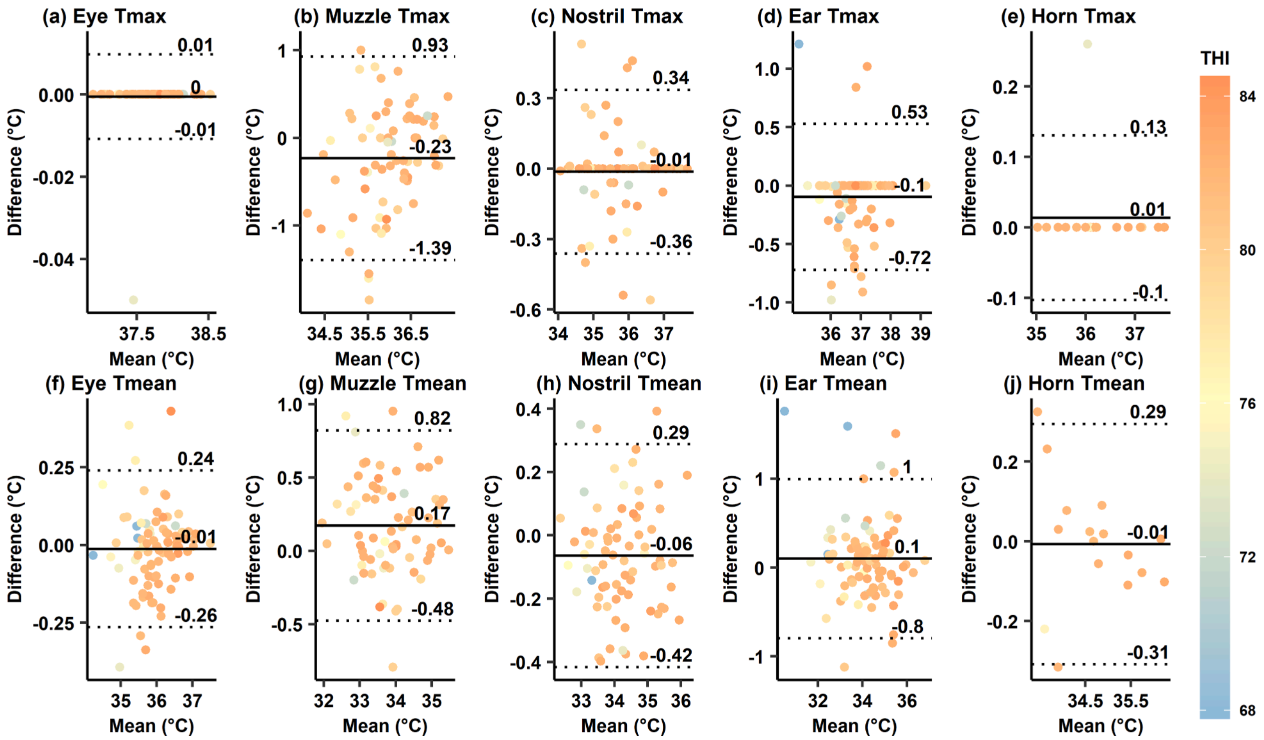

Fig. 10. Bland-Altman plots showing the agreement between manual collection and ground-truth annotation on the testing set (n = 93) in terms of the maximum and mean temperatures (Tmax and Tmean) of five facial landmarks. The solid and dashed lines represent mean difference and 95 % limits of agreement, respectively. Datapoints are coloured by temperature-humidity index (THI).

irregularly shaped landmarks (e.g., ears being seriously influenced by ear tags). Thus, the common practice of using manual collection in relevant studies may obtain under- or over-estimated temperatures.

Other studies compared automated measured and manually collected temperatures. For example, Lowe et al. (2020) reported an average difference between automated measured and manually collected maximum eye temperature of 0 ± 0.001 °C. Wang et al., (2022a) obtained an average difference of 0.051 °C and 0.042 °C between the automated measured and manually collected temperatures of the left and right maximum eye temperatures, respectively. We did not compare automated measurements with manual collections, and thus cannot compare our results with theirs. As discussed before, manually collected maximum eye temperature can be a good proxy of ground truth. Therefore, it is sensible to admit that the present study, as well as the abovementioned studies, all work for automated maximum eye temperature collection. However, our proposed semantic segmentation model can provide more valuable information about mean temperatures which have been determined to more appropriately reflect core body temperatures (Yan et al., 2021).

Besides, the differences between the proposed method and the ground truth (Fig. 9) as well as between the traditional method and the ground truth (Fig. 10) show a homogeneous distribution around their mean differences in most cases, indicating no visible proportional error of one method versus the other. A THI-related colour code was added to confirm whether ambient environments had affected the agreement between the results and their ground truth. It is obvious that the temperature difference stayed homogeneous over the THI range we observed, demonstrating the good robustness of the proposed method against extreme thermal conditions.

3.6. Limitations and future work

Infrared images taken under a thermoneutral environment were limited since the data was originally collected for a heat stress study. Thus, our method should be prioritised for studies in heat stress evaluation. For example, automated heat stress recognition can be achieved by inputting our segmentation results into a deep learning-based thermal level classification (Pacheco et al., 2022). For applications under a thermally comfortable situation, we speculate that our model would remain robust since the larger temperature difference between the animals and their environments in such a situation should increase the separability of ROIs in the infrared images. However, further studies should be conducted in which more data from thermoneutral environments are collected to validate and improve the generalisability of the proposed network.

It should also be noted that the images used for modelling were taken at a fixed distance (i.e., 1 to 1.5 m) from the cows to produce a consistent size of cattle faces. This distance is aligned with most previous studies using IRT to measure bovine body surface temperatures (Montanholi et al., 2015; Peng et al., 2019). However, it may be difficult to reach such a close distance to the cows in practice, except for specific locations like feeding stations (Lowe et al., 2020) and the entry to the milking parlour (Zhang et al., 2020). Direct application of the proposed network to a real-world situation with a greater distance between the cows and the camera should result in a lower segmentation accuracy since facial landmarks in the image taken from a greater distance will be represented by a lower number of pixels.

Plus, the angle of view between the camera and the object is known to influence the infrared emissivity and thus affect the IRT temperature (Muniz et al., 2015). Although we have successfully segmented specific facial landmarks in infrared images taken from different camera angles to the cows, the variation in the angle of view would definitely affect the imaging temperature. Temperature correction is not the focus of this study but should be investigated in future studies in order to obtain reliable temperature readings in more challenging practical cases. A recent study is a promising step forward, in which a response surface method was developed for correcting the effect of distance and the angle of view on the IRT temperature of pigs (Wang et al., 2023).

In addition, many other parameters would have an impact on the results of IRT, such as coat colour, sunlight exposure, emissivity, and the resolution and accuracy of the camera. Thus, the proposed tool may not be directly applied to infrared images acquired in different settings of these parameters. These gaps highlight the need for additional data collection targeted for more complex and practical conditions.

4. Conclusions

Collectively, our work provides relevant studies with an automated tool for collecting facial temperature from cattle infrared images. This method is robust against usual interfering factors including camera angle and extreme ambient environment. However, additional training with data supplemented on a variety of influencing factors (e.g., the distance between the camera and the cows, coat colour) and temperature correction against these factors are required before it can be integrated into on-farm automated monitoring of animal health and welfare.

CRediT authorship contribution statement

Hang Shu: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft. **Kaiwen Wang:** Methodology, Software, Investigation, Data curation, Writing – review & editing. **Leifeng Guo:** Conceptualization, Resources, Project administration, Funding acquisition. **Jérôme Bindelle:** Conceptualization, Methodology, Writing – review & editing. **Wensheng Wang:** Conceptualization, Resources, Supervision, Funding acquisition.

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.

Data availability

We have included the link to our data in the manuscript (Data availability)

Acknowledgements

The first author (H. Shu) was funded by a PhD grant from China Scholarship Council [202203250080]. This work was supported by the Key Research and Development Program of Hebei Province [22326609D] and the Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences [CAAS-ASTIP-2016-AII]. The authors are grateful to Yinxiang dairy farm for their assistance in data collection.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2024.108614.

References

Altman, D.G., Bland, J.M., 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. Journal of the Royal Statistical Society. Series D (the Statistician) 32, 307–317.

Burhans, W.S., Rossiter Burhans, C.A., Baumgard, L.H., 2022. Invited review: Lethal heat stress: The putative pathophysiology of a deadly disorder in dairy cattle. J. Dairy Sci. 105, 3716–3735.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, the European conference on computer vision, pp. 801-818.

Chen, X., Ogdahl, W., Hanna, L., Dahlen, C., Riley, D., Wagner, S., Berg, E., Sun, X., 2021. Evaluation of beef cattle temperament by eye temperature using infrared thermography technology. Comput. Electron. Agric. 188, 106321.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, the IEEE conference on computer vision and pattern recognition, pp. 1251-1258.

Chu, M., Li, Q., Wang, Y., Zeng, X., Si, Y., Liu, G., 2023. Fusion of udder temperature and size features for the automatic detection of dairy cow mastitis using deep learning. Comput. Electron. Agric. 212, 108131.

Collier, R.J., Laun, W.H., Rungruang, S., Zimbleman, R.B., 2012. Quantifying Heat Stress and Its Impact on Metabolism and Performance, Florida Ruminant Nutrition Symposium. University of Florida, Gainesville, FL, USA, pp. 74–83.

Cuthbertson, H., Tarr, G., González, L., 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.

Cuthbertson, H., Tarr, G., Loudon, K., Lomax, S., White, P., McGreevy, P., Polkinghorne, R., González, L.A., 2020. Using infrared thermography on farm of origin to predict meat quality and physiological response in cattle (Bos Taurus) exposed to transport and marketing. Meat Science 169, 108173.

Deng, J., Dong, W., Socher, R., Li, L.J., Kai, L., Li, F.-F., 2009. In: ImageNet: A Large-Scale Hierarchical Image Database, pp. 248–255.

Gloster, J., Ebert, K., Gubbins, S., Bashiruddin, J., Paton, D.J., 2011. Normal variation in thermal radiated temperature in cattle: implications for foot-and-mouth disease detection. BMC Vet. Res. 7, 73.

Halachmi, I., Guarino, M., 2016. Precision livestock farming: a 'per animal'approach using advanced monitoring technologies. Animal 10, 1482–1483.

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., Xu, C., 2020. Ghostnet: More features from cheap operations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580–1589.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hoffmann, G., Schmidt, M., Ammon, C., Rose-Meierhöfer, S., Burfeind, O., Heuwieser, W., Berg, W., 2013. Monitoring the body temperature of cows and calves using video recordings from an infrared thermography camera. Vet. Res. Commun. 37, 91–99.

Jaddoa, M.A., Gonzalez, L., Cuthbertson, H., Al-Jumaily, A., 2021. Multiview Eye Localisation to Measure Cattle Body Temperature Based on Automated Thermal Image Processing and Computer Vision. Infrared Physics and Technology, p. 119.

Jorquera-Chavez, M., Fuentes, S., Dunshea, F.R., Warner, R.D., Poblete, T., Jongman, E. C., 2019. Modelling and Validation of Computer Vision Techniques to Assess Heart Rate, Eye Temperature, Ear-Base Temperature and Respiration Rate in Cattle. Animals 9, 1089.

Khan, M., El Saddik, A., Alotaibi, F.S., Pham, N.T., 2023a. AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network. Knowledge-Based Systems 270, 110525.

Khan, M., Saeed, M., Saddik, A.E., Gueaieb, W., 2023b. ARTriViT: Automatic Face Recognition System Using ViT-Based Siamese Neural Networks with a Triplet Loss. In: 2023 IEEE 32nd International Symposium on Industrial Electronics (ISIE), pp. 1–6.

Kim, S., Hidaka, Y., 2021. Breathing Pattern Analysis in Cattle Using Infrared Thermography and Computer Vision. Animals 11, 207.

Kütük, Z., Algan, G., 2022. Semantic segmentation for thermal images: A comparative survey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 286–295.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lin, J., Yang, H., Chen, D., Zeng, M., Wen, F., Yuan, L., 2019. Face parsing with roi tanh-warping. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5654–5663.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, the IEEE conference on computer vision and pattern recognition, pp. 3431-3440.

Lowe, G., Sutherland, M., Waas, J., Schaefer, A., Cox, N., Stewart, M., 2019. Infrared Thermography-A Non-Invasive Method of Measuring Respiration Rate in Calves. Animals 9.

Lowe, G., McCane, B., Sutherland, M., Waas, J., Schaefer, A., Cox, N., Stewart, M., 2020. Automated Collection and Analysis of Infrared Thermograms for Measuring Eye and Cheek Temperatures in Calves. Animals 10, 292.

Ma, S., Yao, Q., Masuda, T., Higaki, S., Yoshioka, K., Arai, S., Takamatsu, S., Itoh, T., 2021. Development of Noncontact Body Temperature Monitoring and Prediction System for Livestock Cattle. IEEE Sensors Journal 21, 9367–9376.

Montanholi, Y.R., Swanson, K.C., Schenkel, F.S., McBride, B.W., Caldwell, T.R., Miller, S. P., 2009. On the determination of residual feed intake and associations of infrared thermography with efficiency and ultrasound traits in beef bulls. Livestock Science 125, 22–30.

Montanholi, Y.R., Lim, M., Macdonald, A., Smith, B.A., Goldhawk, C., Schwartzkopf-Genswein, K., Miller, S.P., 2015. Technological, environmental and biological factors: referent variance values for infrared imaging of the bovine. Journal of Animal Science and Biotechnology 6, 27.

Muniz, P.R., Magalhães, R.d.S., Cani, S.P.N., Donadel, C.B., 2015. Non-contact measurement of angle of view between the inspected surface and the thermal imager. Infrared Physics & Technology 72, 77-83.

NRC, 1971. A Guide to Environmental Research on Animals. National Academy Press, Washington, DC, USA, p. 374.

Pacheco, V.M., Sousa, R.V., Sardinha, E.J.S., Rodrigues, A.V.S., Brown-Brandl, T.M., Martello, L.S., 2022. Deep learning-based model classifies thermal conditions in dairy cows using infrared thermography. Biosys. Eng. 221, 154–163.

Peng, D., Chen, S., Li, G., Chen, J., Wang, J., Gu, X., 2019. Infrared thermography measured body surface temperature and its relationship with rectal temperature in dairy cows under different temperature-humidity indexes. Int. J. Biometeorol. 63, 327–336.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4510-4520.

Schaefer, A.L., Cook, N.J., Bench, C., Chabot, J.B., Colyn, J., Liu, T., Okine, E.K., Stewart, M., Webster, J.R., 2012. The non-invasive and automated detection of bovine respiratory disease onset in receiver calves using infrared thermography. Res. Vet. Sci. 93, 928–935.

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

Tan, J.-H., Ng, E.Y.K., Rajendra Acharya, U., Chee, C., 2009. Infrared thermography on ocular surface temperature: A review. Infrared Physics & Technology 52, 97–108.

Uddin, J., Phillips, C.J.C., Auboeuf, M., McNeill, D.M., 2021. Relationships between body temperatures and behaviours in lactating dairy cows. Appl. Anim. Behav. Sci. 241, 105359.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534-11542.

Wang, X., Hu, F., Yang, R., Wang, K., 2023. An Infrared Temperature Correction Method for the Skin Temperature of Pigs in Infrared Images. Agriculture 13, 520.

Wang, Y., Kang, X., Chu, M., Liu, G., 2022a. Deep learning-based automatic dairy cow ocular surface temperature detection from thermal images. Comput. Electron. Agric. 202, 107429.

Wang, Y., Kang, X., He, Z., Feng, Y., Liu, G., 2022b. Accurate detection of dairy cow mastitis with deep learning technology: a new and comprehensive detection method based on infrared thermal images. Animal 16, 100646.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Yan, G., Shi, Z., Li, H., 2021. Critical Temperature-Humidity Index Thresholds Based on Surface Temperature for Lactating Dairy Cows in a Temperate Climate. Agriculture 11, 970.

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.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881-2890.

Zheng, Z., Hu, Y., Guo, T., Qiao, Y., He, Y., Zhang, Y., Huang, Y., 2023. AGHRNet: An attention ghost-HRNet for confirmation of catch-and-shake locations in jujube fruits vibration harvesting. Comput. Electron. Agric. 210, 107921.

Zou, K., Chen, X., Wang, Y., Zhang, C., Zhang, F., 2021. A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field. Comput. Electron. Agric. 187, 106242.