Information Processing in Agriculture 11 (2024) 512–523

Contents lists available at ScienceDirect

Information Processing in Agriculture

journal homepage: www.keaipublishing.com/en/journals/information-processing-in-agriculture/

Fusion of RetinaFace and improved FaceNet for individual cow identification in natural scenes

Lingling Yang, Xingshi Xu, Jizheng Zhao, Huaibo Song *

College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China
Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Yangling, Shaanxi 712100, China
Shaanxi Key Laboratory of Agricultural Information Perception and Intelligent Services, Yangling 712100, China

ARTICLE INFO

Keywords:
Cow face recognition
RetinaFace
FaceNet
Deep learning

ABSTRACT

Cows' posture change is the fatal influencing factor for accurate identification of individual cows. To achieve non-contact, high-precision detection and identification of individual cows in farm environment, a cow individual identification method by the fusion of RetinaFace and improved FaceNet was proposed. MobileNet-enhanced RetinaFace was applied to ameliorate the impact of output channel quantity and convolution kernel dynamics using depthwise convolution combined with pointwise convolution. Regression predictions of bovine facial features and keypoints were generated under varying distances, scales and sizes. FaceNet's core feature network was enhanced through MobileNet integration, and the loss function was jointly optimized with Cross Entropy Loss and Triplet Loss to achieve a quicker and more stable convergence curve. The distances between the generated embedding vectors of cow facial features were corresponding to the similarity between cow faces, enabling accurate matching. RetinaFace exhibited detection false negative rates of 2.67%, 0.66%, 2.67%, and 3.33% under conditions of occlusion, no occlusion, low light, and bright light for cow facial detection. For cow facial pattern detection, the false negative rates for black and white patterns, pure black and pure white were 1.33%, 6.00% and 8.00%, respectively. Regarding cow facial posture changes, the false negative rates for face upward, bowing down, profile, and normal posture were 1.33%, 1.33%, 4.00% and 0.66%, respectively. Improved FaceNet model achieved an accuray of 99.50% on training set and 83.60% on test set. In comparison to YOLOX, the recognition model presented in this research demonstrated increased accuracy in cow facial detection under occlusion, no occlusion and strong lighting conditions by 2.67%, 0.40%, and 0.40%, respectively. Moreover, the accuracy for patterns with pure black and pure white tones surpassed that of YOLOX by 1.06% and 5.71%, correspondingly. Additionally, the accuracy rates for face upward, bowing down, profile and normal posture were higher than YOLOX by 2.00%, 3.34%, 2.66% and 0.40%, respectively. The proposed model demonstrates the proficiency in accurately identifying individual cows in natural scenes.

1. Introduction

Automatic identification of dairy cows is essential for modern breeding and management [1–4]. Traditional identification methods are mainly carried out manually and time-consuming [5]. Face recognition [6–7] with machine vision technology achieves contactless and high-precision individual identification in farm environment [8–10]. Especially, individual cow identification in natural scenes plays a fundamental role for smart management of dairy farm. Recently, deep learning algorithms realize high efficiency and accuracy individual cow identification [11], providing potential foundation for individual cow identification on mobile devices.

Multiple methods are proposed for cow identification, including ear grooves, ear lines and soldering iron. In addition, ear tags [12–13] and iris [14] are able to be used for recognition. At present, the most frequently used dairy cow identification method in dairy farms is ear tag marking method, and individual cow number is read manually. Radio frequency identification (RFID) ear tag is developed to identify cow automatically. However, RFID-based cow identification works in limited distance. The RFID reading devices are installed in fixed place, for

Abbreviations: RFID, Radio frequency identification; AP, Average Precision; mAP, mean Average Precision; SSH, Single Stage Headless; BN, Batch Normalization; ReLU, Rectified Linear Unit.

* Corresponding author at: College of Mechanical and Electronic Engineering, Northwest A&F University, Yangling, Shaanxi 712100, China.
E-mail address: songhuaibo@nwsuaf.edu.cn (H. Song).

https://doi.org/10.1016/j.inpa.2023.09.001
Received 16 November 2022; Received in revised form 31 August 2023; Accepted 1 September 2023
Available online 2 September 2023

example, in milking hall. Zin et al. [15] proposed a tracking system for individual cows using an ear tag visual analysis. The proposed system achieved an accuracy of 100.00% for head detection and 92.50% for ear tag digit recognition. Edwards et al. [16] investigated the damage caused by ear tags. In terms of animal welfare, the use of ear tags caused harm to animals. Li [17] pointed out the QR code of the ear tags were ambiguous, making it difficult to scan. Furthermore, ear tags will fall off due to animals' biting and scratching. This situation made it difficult to identify ear tags. Zhao et al. [18] pointed out the traditional ear tags drop due to ear tear or enlarged earhole. The dropping of ear tags made it difficult to read animal's identity. Thus, the development of cow identification methods is still a challenging issue should be solved for smart animal husbandry. Accumulating machine vision [19] methods were proposed to identify cows. Hu [20] employed KPCA-KFDA fusion algorithm to extract nasal striation feature points in cow face, and the accuracy was 95.00%. Wei [21] employed the iris recognition algorithm to extract global and local features of cow's eyes. Even if the iris data was incomplete, the successful recognition rate was 94.70%. However, the process of obtaining the cow iris image is relatively complicated. With the recent development of artificial intelligent technology, deep learning has been documented to be potential method in cow identification and management. Li et al. [22] proposed a dairy cow individual identification method based on improved Mask R-CNN, the Average Precision (AP) was increased by 3.28%, the IoUMask was increased by 5.92%. He et al. [23] proposed a method for individual identification of milking cows. Redmon et al. improved the YOLOv3 [24] deep convolutional neural network, and the recognition accuracy of the improved YOLOv3 model was achieved at 95.91%. Yang et al. [25] introduced coordinate attention mechanism and coordinate convolution module in YOLOv4 [26] network. They proposed a cow facial recognition model fusing coordinate information, the mean Average Precision (mAP) was improved by 0.89%. The above methods of cow identification using deep learning all achieveed high-precision identification. However, when cows community is altered or expanded, the network needs to be retrained on face images set of the whole cows.

Recently, RetinaFace [27] has been proposed to conduct face detection and localization. It utilizes multi-task learning to perform pixel-based face localization on a variety of face scales. The multi-task is trained with joint supervision and self-supervision mechanism. When it is run on image set of the WIDENR FACE [28] hard test rig, RetinaFace outperforms the existing methods with AP improvement of 1.10%. RetinaFace employs lightweight backbone network that it is able to run on a single CPU core in real time. In the field of face recognition, FaceNet [29] algorithm maps the face image directly to Euclidean space and employs the the spatial distance to represent the similarity of the face. Training of FaceNet algorithm requires a small amount of processing on the face image. It achieves accuracy of 99.63% on the LFW [30] dataset and 95.12% on YouTube Faces DB dataset. However, FaceNet uses Inception-ResNetV1 as the backbone feature extraction network. The parameter of the model is 6.8 M, and the calculation is 1550 M times. High model complexity is not suitable for mobile terminals. MobileNet [31] is used as the main feature extraction network for FaceNet. Cross Entropy Loss was used to assist Triplet Loss for convergence and speed up convergence.

The main objective of the research was to develop a detection and recognition algorithm for an intelligent cow face recognition system. The ultimate goal was to accurately identify individuals in farm environment. The specific objectives were:

(1) to develop a dataset of cow face images including variations in pattern, pose, shading and illumination.
(2) to develop a deep learning neural network with fusing of RetinaFace and FaceNet model, improving the model's backbone network and loss function to accurately detect and identify individual cows.

2. Dataset preparation

2.1. Acquisition of dairy cow face images

A diverse dataset consisting of 110 cows were collected from two different dairy farms. Dairy cow face images were sampled in dairy farm in Yangling District, Xianyang City, and Wuzhong, Ningxia Hui Autonomous Region. The data collection encompassed various scenarios commonly encountered in dairy farms, capturing cow facial images in natural farm settings as well as controlled environments. Data collection was conducted both indoors and outdoors, with careful attention to lighting conditions and background clutter. The images were captured from 17 to 21 February 2022, during the hours of 7:00 am-12:00 pm each day. The images were took with a Huawei Nova9 mobile phone. Each image was 1080 × 2340 pixels. Data collection process involved approaching the cows in a non-intrusive manner and capturing facial images at a suitable distance. The cows were accustomed to human presence, ensuring minimal stress during the image capture. The distance between camera and cow was among 60–80 cm. A total of 2,376 face images from 110 cows were obtained in the current study (Table 1).

The current study included four typical situations for cow face detection and identification, including variations in posture, masking, lighting and face color. When the posture altered, the shapes of organs changed sharply. Fig. 1 shows images of cow head raising, lowering and facing sideways. When the image of the cow face was occluded, the texture features of cow face images changed greatly. Fig. 2 shows facial occlusion that caused loss of facial features. Fig. 3 shows cow face images that were captured in different light conditions. Fig. 4 shows cow's face pattern of pure black or white.

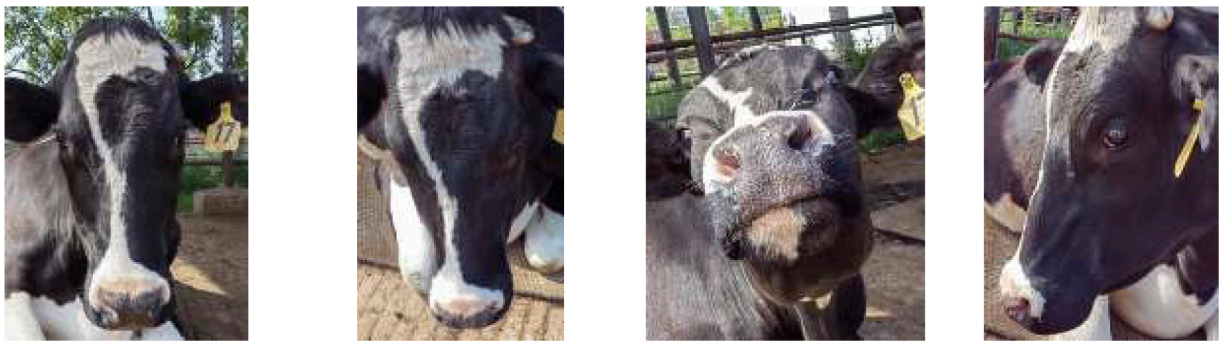
2.2. RetinaFace model dataset preparation

WIDER FACE is the detection dataset of RetinaFace model. The dataset contains the information of face posture, expression, scale, occlusion and illumination. The dataset is divided into 60 categories. The label.txt file includes the image file name, the number of faces in the image, x_1, y_1, *w, h, blur, expression, illumination, invalid, occlusion, pose* and other information. The *blur* represents the blurriness of the image. The *expression* represents the normality of the expression. The *illumination* represents exposure and *invalid* represents occlusion. As shown in Fig. 5, x1, y1, *w, h* represent the coordinates of the upper left corner, width and height of the label box.

In order to reproduce the format of dataset, the size of cow's face image was resized to 4096 × 3072 pixels. As shown in Fig. 5, the cow image was labeled with a rectangular frame and five key points using the Labelme software. The format of the marked file was.json. Then, the.json file was converted to a.txt file containing the position of the cow face image, the number of labeled borders, the coordinates and properties of

Table 1
Cow face image dataset.

Name	Classify	Quantity	Proportion (%)
Face pattern style	Black and white	1732	72.90
	Pure black	504	21.21
	Pure white	140	5.89
Posture change	Face upward	500	21.04
	Bow one's head	405	17.05
	Profile	538	22.64
	Normal	933	39.27
Shelter	Sheltered	792	33.33
	Unobstructed	1584	66.67
Number of dairy cow faces in one image	Single cow	1752	73.73
	Multiple cows	624	26.26
Illumination change	Low light	1173	49.37
	Bright light	1203	50.63

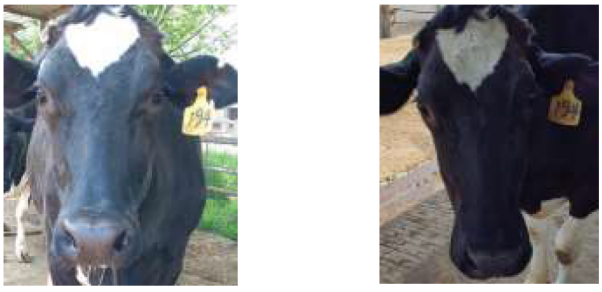


(a) Positive image of cow (b) Cow head down image (c) Cow head up image (d) Cow side face image

Fig. 1. Different postures of cow's face.

(a) Cow unobstructed image (b) Cows have occluded images

Fig. 2. Comparison image of cow face with occlusion.



(a) Image of cow face in bright light (b) Image of cow face in low light

Fig. 3. Cow's face under different illumination.

(a) Image of pure black cow face (b) Image of pure white cow face

Fig. 4. Face of cows with different color pattern.

the cow face border.

2.3. FaceNet model dataset preparation

The deep learning models RetinaFace had been used for cow face detection. Three predicted results were obtained. The prediction results were classification prediction, face frame and bull face key point respectively. The prediction results of key points included left eye, right eye, left nostril boundary point, right nostril boundary point and lower nostril center point. The coordinates of the eyes of the cow's face were obtained. When cow face was successfully detected by RetinaFace, the image was included to further analysis. In order to make FaceNet dataset, this research conducted alignment correction on images. Alignment correction was beneficial to the cow's facial feature extraction. In this research, the correction operation was carried out by rotating the binocular coordinates of cows. As shown in Fig. 6(a), inclination angle θ represents the angle between the line A of both eyes and the horizontal line B. M represents the center point of the image. As shown in Fig. 6(b), it was the cow's face image after the alignment correction operation.

If the coordinates of the left eye of the cow image were (x_1, y_1) and the coordinates of the right eye of the cow image were (x_2, y_2), the calculation of the rotation angle θ was shown in Eqs.(1)-(3):

$$\theta = arctan\frac{y}{x} \times \frac{180}{\pi} \tag{1}$$

$$y = y_1 - y_2 \tag{2}$$

$$x = x_1 - x_2 \tag{3}$$

3. Network structure analysis

3.1. RetinaFace network model

Based on the RetinaNet structure, RetinaFace used a feature pyramid technique to achieve multi-scale information fusion. It was designed for multi-task learning with a combination of extra supervision and self-supervision. RetinaFace was added with a self-supervised network decoder branch, different sized faces can be positioned at the pixel level. The network structure was shown in Fig. 7.

RetinaFace was designed to use MobileNetV1-0.25 or ResNet as the backbone feature extraction network for training. After training the two backbone networks, MobileNetV1-0.25 was used as the backbone feature network. This network was developed by using of deeply separable convolutional. The depth separable convolution adopted the combination of channel by channel convolution and point by point convolution. The effect between the number of output channels and the convolution kernel was eliminated. The comparison diagram of deep separable convolution and standard convolution was shown in Fig. 8

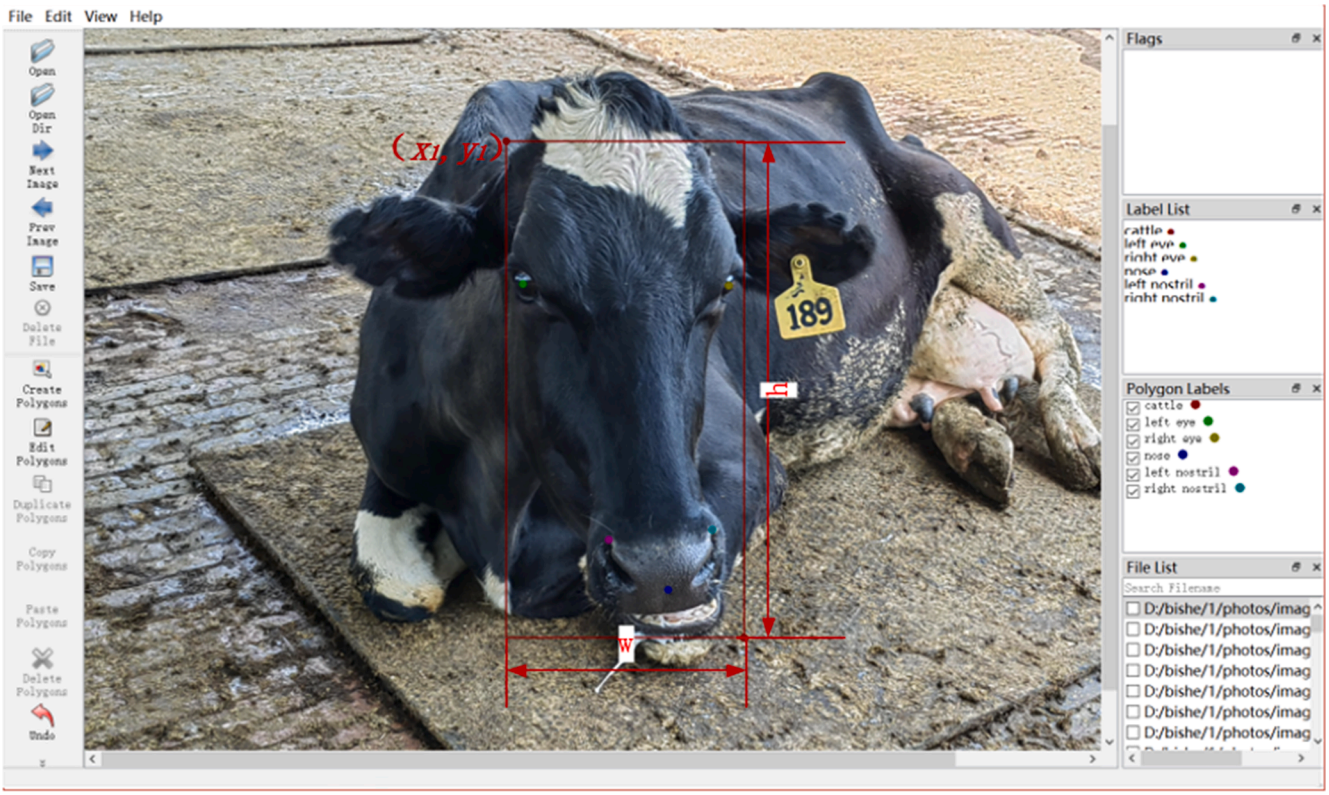
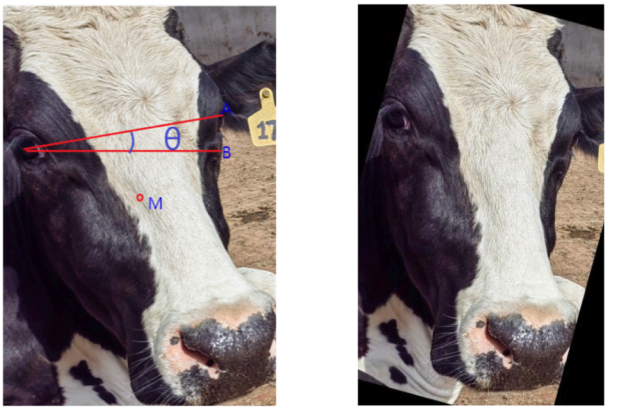


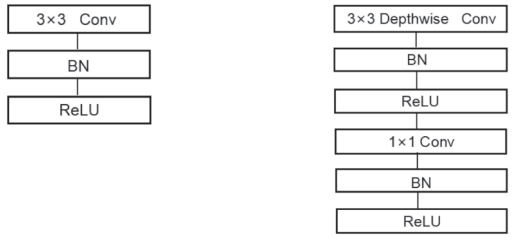
Fig. 5. Annotation drawing of cow image.



(a) Original cow image (b) Crop aligned image

Fig. 6. Dairy cow face image correction map.

[31]. Compared with the standard convolution, the depth separable convolution parameter was reduced to 14.20% of the normal convolution.



(a) Standard convolutionimage (b) Depth separableconvolution image

Fig. 8. Comparative plot of marked vertebral convolution and deep convolution.

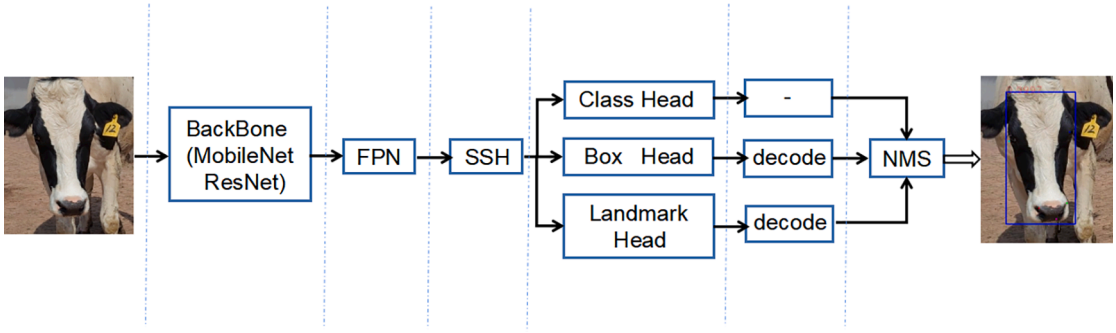


Fig. 7. RetinaFace network structure diagram.

Multiscale problems in object detection were solved by feature gin-tower networks. By changing the network connection, the model was able to improve the small object detection performance without increasing the model computation. Feature pyramid layer from P2 to P6 adopted by RetinaFace. P2 to P5 were designed to output results from ResNet residual stage using top-down and horizontal connections. After obtaining three effective layers, Single Stage Headless (SSH) was used to enhance feature extraction. 5 × 5 and 7 × 7 convolution were replaced by 3 × 3 stacked convolution, thus three effective feature layers were obtained. The regression results of classification, frame and facial key points were obtained through these three effective feature layers.

The loss [27] function of the RetinaFace was shown in Eq.(4):

$$L = L_{cls}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) + \lambda_3 p_i^* L_{pixel} \tag{4}$$

where, in the classification loss function $L_{cls}(p_i, p_i^*)$, p_i was the prediction probability of the cow's face. When$p_i = 1$, it is positive anchor, and$p_i = 0$, it is negative anchor. In the Regression loss function $L_{box}(t_i, t_i^*)$, t_i={t_x, t_y, t_w, t_h} represents the prediction frame coordinates related to the positive anchor. t_i^*={t_x^*, t_y^*, t_w^*, t_h^*} represents the real frame coordinates related to the positive anchor. In the Regression loss function $L_{pts}(l_i, l_i^*)$, l_i={l_{x1}, l_{y1}, …l_{x5}, l_{y5}} represents the five key points of the cow's face. l_i^*={l_{x1}^*, l_{y1}^*, …l_{x5}^*, l_{y5}^*} represents the five reference points of cow's face.

3.2. FaceNet network model

The input image was converted into the feature vector in the European space by the FaceNet model. The European distance between vectors was used to measure the similarity between the input images. The threshold was used to determine whether the two face images belong to the same individual. FaceNet had an accuracy of 99.63% on the LFW dataset and 95.12% on the YouTube faces DB dataset. The network structure of FaceNet was shown in Fig. 9. As shown in Fig. 9, Batch represented the image sample that the detected cow face was cut to a fixed size. Deep architecture represented the adoption of a deep learning framework. Embedding represented the feature vector after feature normalization. Embedding was used to embed a high-dimensional space of all words into a continuous vector space of much lower dimensionality. Each word or phrase was mapped as a vector over a real number field. In PyTorch, there was a dedicated layer for word vectors Embedding. It was used to implement the mapping of words to word vectors.

The loss function Triplet Loss was proposed in the FaceNet model. For the cow face image ×,the Triplet Loss represented that $f(x)$ was measured within the Euclidean distance. By using the Triplet Loss, the distance vector of the same cow face image became smaller, the distance vector of different cow face images became larger. The loss [29] function was shown in Eq.(5):

$$L = \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{5}$$

where, a represents Anchor, the 128 dimensional face feature vector obtained from the reference image; p represents positive, the 128 dimensional cow face feature vector obtained from the same cow face image as the reference image; n represents negative, the 128 dimensional cow face feature vector obtained from the same cow face image as the reference image. The training process of the network was shown in Fig. 10. Through continuous learning and training, the distance between the anchor and the positive was smaller, and the distance between the anchor and the negative became farther.

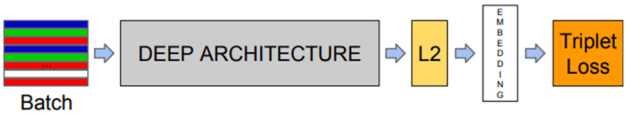


Fig. 9. Network structure diagram of FaceNet model.

The training recognition results were shown in Fig. 11. When the two input images belonged to the same cow, the distance recognized by FaceNet network became smaller. When the two input images belonged to the different cow, the distance recognized by FaceNet network became farther.

In the process of FaceNet recognition, the face features of each dairy cow were acquired in real time. At the same time, the features of the input cow face image were compared with the faces of all cows in the database. The eigenvector distance between the input image and the face image in the database were calculated. The most similar cow serial number in the database was obtained. If the distance of the serial number was less than the threshold, the recognition was considered successful. The recognition process of FaceNet was shown in Fig. 12.

In this research, Inception-ResNetV1 was selected as the backbone feature extraction network of FaceNet. It had a large number of parameters and calculation. At the same time, Triplet Loss was applied to converge the network, and the network convergence speed was slow and difficult. In this research, the FaceNet model was improved from two aspects:

1) Inception-ResNetV1 was used as the backbone feature extraction network in the original FaceNet, this approach failed to conduct recognition on mobile. In this research, lightweight neural network MobileNetV1 was used as the FaceNet backbone feature extraction network.
2) Ternary loss was used in the original FaceNet. Triplet Loss was slow and difficult to calculate. In this research, Cross Entropy Loss was used to assist Triplet Loss to converge.

3.3. Improved FaceNet network model

3.3.1. Improved the backbone network

FaceNet used Inception-ResNetV1 as the backbone feature extraction network. The parameter of the model was 6.8 M, and the calculation was 1550 M times. High model complexity was not suitable for mobile terminals. MobileNet, a lightweight convolutional neural network, had smaller volume, less calculation and high precision. We utilized MobileNet as main feature extraction network for FaceNet. After introducing the depth separable convolution, many application functions was able to be realized on mobile device. As shown in Fig. 13, it was a convolution process of deep separable convolution. Ordinary convolution was replaced by deep convolution and point convolution. Except for the first convolution layer and the last fully connected layer, all convolution layers were connected by (Batch Normalization) BN and (Rectified Linear Unit) ReLU.

The deep separable convolution was shown in Eq.(6):

$$G_{k,l,m} = \sum_{i,j} K_{i,j,m} \bullet F_{k+i-1,l+j-1,m} \tag{6}$$

where, K represents the convolution kernel, F represents the features to be convoluted, i, j represents the position of the pixel, M represents the M^{th} channel of a convolution kernel, N represents the number of convolution kernels.

The number of deep separable convolution output channels and the size of convolution kernel were large. Compared with the standard model, the overall calculation power consumption was reduced more. Depth separable convolution was able to maintain the accuracy of image recognition while reducing the complexity of network model. Main channel convolution and point convolution were adopted by MobileNet, the number of parameters and the volume of calculations had been reduced. In this research, MobileNetV1 was used as the backbone feature extraction network.

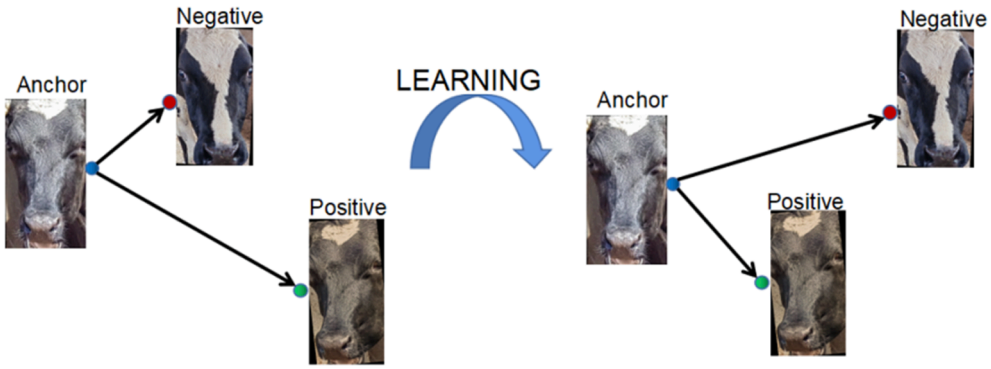


Fig. 10. FaceNet model network learning training chart.

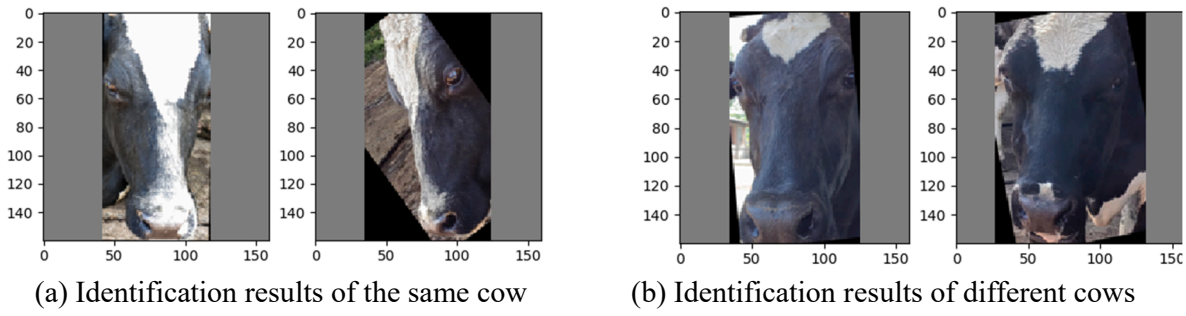


Fig. 11. FaceNet network identification result chart.

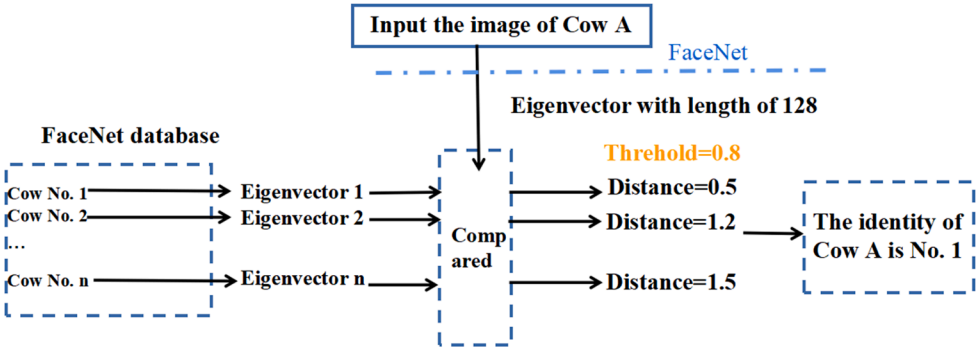


Fig. 12. FaceNet model network identification process diagram.

Fig. 13. MobileNet convolution process diagram.

3.3.2. Improved loss function

In this research, Cross Entropy Loss and Triplet Loss were used as the overall loss. Cross-Entropy measures the difference between two different probability distributions in the same random variable. The smaller the distance means that the two probabilities are similar. A larger distance indicates a larger difference between the two probabilities. In machine learning, Cross-Entropy was expressed as the difference between the true probability distribution and the predicted probability distribution. In this way, a classifier was constructed, and the Cross Entropy Loss assist Triplet Loss to converge. For a single sample, suppose the true distribution is y_i, the network output distribution is p_i, the total number of categories is n. The Cross Entropy Loss was shown in Eq.(7):

$$L = -\sum_{i=1}^{n} y_i log(p_i) \tag{7}$$

In Eq. (7), y_i denotes the label of sample i. When it is a positive class, y_i is 1. When it is a negative class, y_i is 0. For a batch, the formula for the Cross Entropy Loss function for the single-label n classification task was

shown in Eq.(8):

$$L = -\frac{1}{batch\_size}\sum_{j=1}^{batch\_size}\sum_{i=1}^{n} y_{ij} log(p_{ij}) \quad (8)$$

In Eq. (8), y_{ij} denotes the label of sample i. When it is a positive class, y_{ij} is 1. When it is a negative class, y_{ij} is 0. When the Cross-Entropy value was smaller, the model prediction effect was better. Cross-Entropy classifier was used for auxiliary training to accelerate convergence speed.

3.4. Experimental platform configuration

The development platform used in this research was PyCharm, and the specific configuration of the platform was shown in Table 2.

4. Experimental results and analysis

4.1. Experimental results and analysis of the RetinaFace model

4.1.1. Training based on RetinaFace model

During training, the Epoch value was set to 815, the batch_size was 8, and the maximum learning rate of the model was 0.01. Backbone extraction networks of MobileNet and ResNet were trained for comparison, the loss function of the experiments was shown in Fig. 14.

As shown in Table 3, the parameters and calculations of MobileNet and ResNet were compared.

The experimental results showed that, with the increase of the number of iterations, although the loss of MobileNetV1 curve was slightly higher than that of ResNet, the number of parameters and computation of the MobileNet model were significantly reduced. Therefore, MobileNetV1 was chosen as the backbone extraction network for recognition training.

4.1.2. Detection results based on the RetinaFace model

The detection results of cow face images in different poses were shown in Fig. 15. As shown in Fig. 15, the RetinaFace model was able to detect different angles of cows' face images well.

The image detection results of different facial patterns were shown in Fig. 16. As shown in Fig. 16, the RetinaFace model was able to detect cow face images with different patterns well.

The results for cow face images with occlusion, illumination changes and multiple cows in one image were shown in Fig. 17. As shown in Fig. 17(a) and (b), the presence of an occluded cow face image can be detected well. As shown in Fig. 17(c) and (d), the images of cow faces were detected under changing light conditions. As shown in Fig. 17(e), there were multiple cows in one image, and cows were undetected in this image.

In this research, cow face images were selected to measure the leakage rate of the model. The selected images included single cow or multiple cows. Selected images of single cows included images of cow faces with different patterns, lighting, shading and posture. The missing rate results of RetinaFace model were shown in Table 4. As shown in Table 4, The undetected rate of solid color cows and multi head cows was high.

Table 2
Experimental environment parameter configuration.

Name	Parameter
Processor	Intel(R) Core(TM) i5-8250U CPU @ 1.60 GHz 1.80 GHz
Graphics card	NVIDIA GeForce MX 150
Graphics platform	CUDA10.2
Programming language	Python3.6
Deep learning framework	PyTorch
System type	64-bit operating system

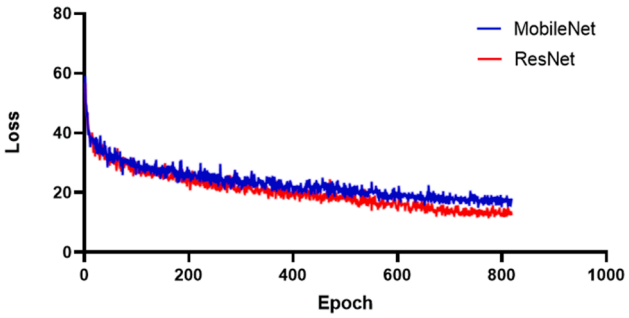


Fig. 14. MobileNet and ResNet network Loss diagram.

Table 3
Comparison table of parameters of different network models.

Models	Number of participants(M)	Calculated volume(M)
MobileNet	4.2	569
ResNet	6.8	1550

4.2. Experimental results and analysis of the improved FaceNet model

For the cow face dataset produced in this research, it was randomly divided into training set and test set according to the ratio of 7:3. In the training period, the number of iterations was set to 250, the batch_size was set to 8 and the maximum learning rate was set to 0.001.

4.2.1. Results of the improved loss function

In this research, the backbone extraction network in the FaceNet was set to Inception-ResNetV1. The loss function was Cross Entropy Loss combined with Triplet Loss. The loss function designed in this research was trained to verify the superiority of the joint loss function. The trained loss curve was shown in Fig. 18.

As shown in Fig. 18, the joint Cross Entropy Loss and Triplet Loss were the overall reduced with iterations increase. The loss function converges faster after Epoch was 5. The joint loss function started to converge stably when the iteration was about 140 times. The Triplet Loss function was used for convergence, and the curve was close to convergence at about 180 iterations. Therefore, the combination use of Cross Entropy Loss and Triplet Loss made the convergence speed faster and the convergence curve more stable.

4.2.2. Backbone network comparative analysis

The backbone feature extraction network of the original FaceNet model was Inception-ResNetV1. The backbone network of the original model had the problems of slow operation speed and high model complexity. In this research, MobileNetV1 was chosen to be the backbone network for recognition training to address the problems. To verify the effectiveness of the improved backbone network, this research set the loss function as Cross Entropy Loss combining Triplet Loss. Then the FaceNet models of Inception ResNetV1 and MobileNetV1 were trained and compared respectively. The training results were shown in Table 5.

When MobileNetV1 was chosen as the backbone extraction network for recognition training, the training accuracy of MobileNetV1 was 99.50%. Its training accuracy was 0.30% lower than that of the Inception-ResNetV1. The Verification accuracy of MobileNetV1 was 83.60%. The Verification accuracy was 0.50% lower than the Inception-ResNetV1. The number of parameters of the MobileNetV1 model was reduced to 61.80% of the Inception-ResNetV1. The computational effort was reduced to 36.70% of Inception-ResNetV1. The size of the MobileNetV1 trained model file was 12.9 MB, which was 14.80% of the Inception-ResNetV1 model. Therefore, MobileNetV1 was selected as the backbone extraction network of FaceNet model. This method was able to reduce the computational complexity of the model while keeping the

Fig. 15. Detection of cows' faces in different postures.

Fig. 16. Different patterned cow face detection map.

almost identical training and testing accuracy. At the same time, it was more convenient to embed this model into the mobile terminal, and cattle identity was more effectively managed by farmers and herdsmen.

4.3. Discussions

4.3.1. Analysis of cow identification results for different pattern colours

Cow face pattern included black and white, pure white and pure black, and it was a key factor in feature extraction and face recognition. When the color of the cow's face pattern was pure white and black, it was difficult to identify the cow. This research selected 300 cow face images to test the recognition of cows with different facial patterns. Among the selected images, the number of single cow face images with black and white, pure black, and pure white were 185, 95, and 35.

As shown in Table 6, it was the recognition results in this research model and YOLOX model. The accuracy of YOLOX model was 0.54% higher than this model on the black and white cow dataset. The accuracy of this model was 1.06% and 5.71% higher than YOLOX model on the dataset of cows with pure black and white patterns. The results showed that that the proposed model was able to extract the deeper features of cow face more effectively, rather than relying solely on the speckle features for classification.

4.3.2. Analysis of cow identification results for different posture change

When the posture changes, the shapes of organs observed from different angles are different. This research selected 700 cow face images to test the recognition of cows with different posture change. Among the selected images, the number of single cow face images with face upward, bow one's head, profile, and normal were 150, 150, 150, and 250. As shown in Table 7, it was the recognition results of cows with different posture change in this research model and YOLOX model.

As shown in Table 7, it was the recognition results in this research model and YOLOX model. The accuracy of this model was 2.00%, 3.34%, 2.66% and 0.40% higher than YOLOX model on the dataset of cows with face upward, bow one's head, profile and normal. The results showed that the model in this research was able to well adapt to the changes of cow's facial posture.

4.3.3. Analysis of cow identification results of cow with shelter

When the image of the cow face was occluded, the texture features of cow face images changed greatly. This research selected 350 cow face images to test the recognition of cow face with occlusion. Among the selected images, the number of single cow face images with face sheltered, and unobstructed were 100, and 250. As shown in Table 7, it was the recognition results of cows with shelter change in this research model and YOLOX model.

As shown in Table 8, it was the recognition results in this research model and YOLOX model. The accuracy of this model was 2.67% and 0.40% higher than YOLOX model on the dataset of cows with Sheltered and Unobstructed. The results showed that the model in this research can better capture the facial information of occluded cow.

4.3.4. Analysis of cow identification results of cow with illumination change

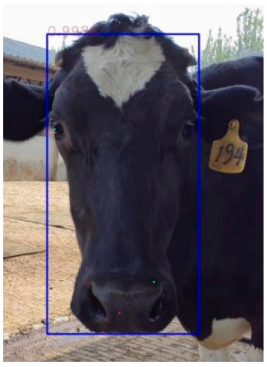
This research selected 300 cow face images to test the recognition of cow face with illumination change. Among the selected images, the number of single cow face images with low light, and bright light were 150, and 150. As shown in Table 9, it was the recognition results in this research model and YOLOX model. The accuracy of this model was 0.40% higher than YOLOX model on the dataset of cows with and bright light. Both models were able to recognize the identity of cows well when the light changes.

As shown in Fig. 19, it was a comparison chart of two model recognition results. It can be seen from the figure that No.0144 pure black cow was wrongly identified by YOLOX model. For No.5 black and white pattern cow, the model in this research can detect and recognize the cow face well. For the sheltered No.10 cattle, the model identification in this research was correct compared with YOLOX model. It showed that the model proposed in this research can identify cows with

(a) Single uncovered cow figure

(b) Single Head with Covered Cow Figure

(c) Image of cow face in bright light



(d) Image of cow face in low light

(e) Picture of multiple cows' faces

Fig. 17. Effect picture of cow face image detection.

Table 4
Comparison table of parameters of different network models.

Name	Classify	Leakage rate (%)	Quantity
Face pattern style	Black and white	1.33	150
	Pure black	6.00	50
	Pure white	8.00	50
Posture change	Face upward	1.33	150
	Bow one's head	1.33	150
	Profile	4.00	150
	Normal	0.66	150
Shelter	Sheltered	2.67	150
	Unobstructed	0.66	150
Number of dairy cow faces in one image	Multiple cows	4.67	150
Illumination change	Low light	2.67	150
	Bright light	3.33	150

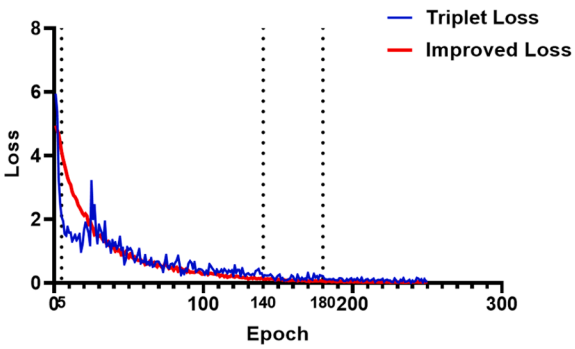


Fig.18. Two loss function loss line graphs.

Table 5
Comparison table of parameters of different network models.

Models	Number of participants (M)	Volume of calculations (M)	Train_accuracy (%)	Val_accuracy (%)
MobileNetV1	4.2	569	99.50	83.60
Inception-ResNetV1	6.8	1550	99.80	84.10

Table 6
Comparison of identification performance for this research and YOLOX in different patterns.

Different classifications	Number of images	Models	Accuracy(%)
Black and White	185	Ours	95.67
		YOLOX	96.21
Pure black	95	Ours	92.63
		YOLOX	91.57
Pure white	35	Ours	88.57
		YOLOX	82.86

Table 7
Comparison of identification performance for this research and YOLOX in posture change.

Different classifications	Number of images	Models	Accuracy(%)
Face upward	150	Ours	97.33
		YOLOX	95.33
Bow one's head	150	Ours	96.67
		YOLOX	93.33
Profile	150	Ours	95.33
		YOLOX	92.67
Normal	250	Ours	98.40
		YOLOX	98.00

Table 8
Comparison of identification performance for this research and YOLOX in cow with shelter.

Different classifications	Number of images	Models	Accuracy(%)
Sheltered	150	Ours	96.67
		YOLOX	94.00
Unobstructed	250	Ours	98.40
		YOLOX	98.00

Table 9
Comparison of identification performance for this research and YOLOX in cow with illumination change.

Different classifications	Number of images	Models	Accuracy(%)
Low light	150	Ours	96.67
		YOLOX	96.67
Bright light	150	Ours	98.40
		YOLOX	98.00

different patterns well.

4.3.5. Video detection and recognition experiments

In order to verify the application of the algorithm model in the field dynamic scene, it was tested in the field collected cow face video. The video frame size of cow face image was 1920 × 1080 pixels, the frame rate was 29.64 fps. The experimental results were shown in Fig. 20.

At video frame numbers 166, 442 and 795, the cow's face was well detected in the image. As shown in Fig. 20(d), the image of the cow's face in the lower left corner of the figure was not detected.

4.3.6. Improvements and contribution

In this research, applying RetinaFace to cow facial detection resulted in regression prediction outcomes for cow facial regions and key points. We used Cross Entropy Loss in conjunction with Triplet Loss function to converge the FaceNet model. The loss function converged faster and the convergence curve was more stable. MobileNetV1 was selected as the backbone extraction network of the FaceNet model. The improved model reduced the computational complexity of the model while maintaining the accuracy of training and testing.

On the pure black and pure white cow dataset, the proposed model achieved better recognition results compared to YOLOX. The results showd that this model can identify cows with different patterns. In comparison to YOLOX, the recognition model presented in this research demonstrated increased accuracy in cow facial detection under occlusion, no occlusion, and strong lighting conditions by 2.67%, 0.40%, and 0.40%, respectively. Moreover, the accuracy for patterns with pure black and pure white tones surpassed that of YOLOX by 1.06% and 5.71%, correspondingly. Additionally, the accuracy rates for face upward, bowing down, profile, and normal posture were higher than YOLOX by 2.00%, 3.34%, 2.66%, and 0.40% respectively. Compared to previous research on the use of deep learning for cow identification, this model could be trained and learned quickly when the cow dataset changes. At the same time, it contributed to the realization of non-contact and high-precision detection and identification of individual cow identity in the dairy farms.

4.3.7. Disadvantages

There were cases of mis-detection when there are multiple cows in the image. As shown in Fig. 21, the first cow in the bottom left corner was mis-detected.

The cow face image in the video was effectively detected and recognized by the algorithm proposed in this study. The recognition result of cow face image was displayed in the lower left corner of the prediction box. From the video frame number 952 in Fig. 20, it could be seen that the detection distortion occurs when the video frame has multiple cows. By analyzing the test results, the proportion of cows with solid patterns in the dataset was low. It was difficult to extract the facial features of cows with solid color patterns. In the future, more solid cows will be collected for training to improve the network detection and recognition capability.

4.4. Conclusions

To achieve high-precision detection and identification of individual cows in farm environment, a cow individual identification method by the fusion of RetinaFace and improved FaceNet was proposed. Applying MobileNet-enhanced RetinaFace to cow facial detection enabled regression prediction of both cow facial regions and key points. FaceNet's core feature network was enhanced through MobileNet integration, and the loss function was jointly optimized with Cross Entropy Loss and Triplet Loss to achieve a quicker and more stable convergence curve. The proposed model adeptly captured subtle distinctions among bovine facial features, achieving a heightened precision in individual identification. This held paramount importance in the domains of farm management and tracking. Experimental results presented indicate that the accuracy for patterns with pure black and pure white tones surpassed that of YOLOX by 1.06% and 5.71%, correspondingly. In comparison to YOLOX, the recognition model presented in this research demonstrated increased accuracy in cow facial detection under occlusion, no occlusion, and strong lighting conditions by 2.67%, 0.40%, and 0.40%, respectively. The facial patterns of cows play a significant role in identification. The dataset pertaining to solid-colored patterned cows was constrained, resulting in an insufficient in-depth analysis of their recognition characteristics. To enhance the research depth, there are plans to gather a more comprehensive dataset in the future.

(a) YOLOX test results

(b) Results of model testing in this research

Fig. 19. Comparison of the recognition results of the two models.

(a) Video frame number: 166 (b)Video frame number: 442

(c) Video frame number: 795 (d) Video frame number: 952

Fig. 20. Recognition effect of cow face image in video.

5. Declaration of generative AI in scientific writing

During the preparation of this work the authors used RetinaFace and FaceNet in order to achieve detection and identification of individual cows. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Ethical statement

The research does not contain any studies conducted by the authors involving human participants. The study was approved by the Medical Ethics Committee (MEC) of NWAFU Experimental Animal Manage Committee of Northwest A&F University.

Funding

This work was supported by the National Natural Science Foundation

Fig. 21. Face detection map of multiple cows.

of China (No. 32272931), and the Shaanxi Provincial Technology Innovation Guidance Planned Program (No. 2022QFY11-02).

CRediT authorship contribution statement

Lingling Yang: Writing – original draft. **Xingshi Xu:** Writing – review & editing. **Jizheng Zhao:** Writing – review & editing. **Huaibo Song:** Writing – review & editing.

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.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 32272931), and the Shaanxi Provincial Technology Innovation Guidance Planned Program (No. 2022QFY11-02). The authors appreciate the funding organization for their financial supports. The authors would like to thank all of the authors cited in this article and the anonymous referees for their helpful comments and suggestions.

References

[1] Yao L, Hu Z, Liu C, Liu H, Kuang Y, et al. Cow face detection and recognition based on automatic feature extraction algorithm. Proceedings of the ACM Turing Celebration Conference - China, Association for Computing Machinery, Chengdu, China 2019.
[2] Kumar S, Tiwari S, Singh SK. Face Recognition of Cattle: Can it be Done? Proc Natl Acad Sci, India, Sect A Phys Sci 2016;86(2):137–48.
[3] Wang Y, Xu X, Wang Z, Li R, Hua Z, Song H. ShuffleNet-Triplet: A lightweight RE-identification network for dairy cows in natural scenes. Comput Electron Agric 2023;205:107632.
[4] Xiao J, Liu G, Wang K, Si Y. Cow identification in free-stall barns based on an improved Mask R-CNN and an SVM. Comput Electron Agric 2022;194:106738.
[5] Lu Y, He X, Wen Y, Wang P. A new cow identification system based on iris analysis and recognition. Int J Biometr 2014;6:18–32.
[6] Xia M, Cai C. Cattle face recognition using sparse representation classifier. ICIC Express Lett Part B: Appl 2012;3:1499–505.
[7] Wang H, Wang Y, Zhou Z, Ji X, Gong D, et al. CosFace: Large Margin Cosine Loss for Deep Face Recognition 2018.
[8] Zin T.T., Phyo C.N., Tin P., Hama H., Kobayashi I. Image Technology based Cow Identification System Using Deep Learning.
[9] Srivastava Y, Murali V, Dubey SR. A Performance Evaluation of Loss Functions for Deep Face Recognition 2020:322–32.
[10] Wang H, Qin J, Hou Q, Gong S. Cattle Face Recognition Method Based on Parameter Transfer and Deep Learning. J Phys Conf Ser 2020;1453(1):012054.
[11] Qiao Y, Su D, Kong He, Sukkarieh S, Lomax S, Clark C. Individual Cattle Identification Using a Deep Learning Based Framework. IFAC-PapersOnLine 2019;52(30):318–23.
[12] Johnston AM, Edwards DS. Welfare implications of identification of cattle by ear tags. Vet Rec 1996;138(25):612–4.
[13] Leslie E, Hernández-Jover M, Newman R, Holyoake P. Assessment of acute pain experienced by piglets from ear tagging, ear notching and intraperitoneal injectable transponders. Appl Anim Behav Sci 2010;127(3-4):86–95.
[14] Gonzales Barron U, Corkery G, Barry B, Butler F, McDonnell K, Ward S. Assessment of retinal recognition technology as a biometric method for sheep identification. Comput Electron Agric 2008;60(2):156–66.
[15] Zin TT, Pwint MZ, Seint PT, Thant S, Misawa S, Sumi K, et al. Automatic Cow Location Tracking System Using Ear Tag Visual Analysis. Sensors 2020;20(12):3564.
[16] Edwards DS, Johnston AM, Pfeiffer DU. A comparison of commonly used ear tags on the ear damage of sheep. Anim Welf 2001;10(2):141–51.
[17] Li J. Problems and Countermeasures of Animal Identification in Product Traceability. The Chinese Livestock and Poultry Breeding 2020;16:29–30.
[18] Zhao Z, Wang F, He S, Liu C, Wang X. Experimental analysis of Tibetan sheep wearing ear tags of different materials and shapes. China Animal Industry 2022:61–2.
[19] Kumar S, Singh SK, Abidi AI, Datta D, Sangaiah AK. Group Sparse Representation Approach for Recognition of Cattle on Muzzle Point Images. Group Sparse Representation Approach for Recognition of Cattle on Muzzle Point Images 2018;46(5):812–37.
[20] Hu Y. Contrast Analysis of Feature Points for Nasal Lines Based on Image Processing. Computer Simulation 2016;33:314–7.
[21] Wei Z. Research on imperfect bovine iris recognition based on the combination of global and local features. 2017.
[22] Li H, Chen G, Pei A. Research on individual recognition of dairy cows based on improved Mask R-CNN. J South China Agric Univ 2020;41:161–8.
[23] He D, Liu J, Xiong H, Lu Z. Individual Identification of Dairy Cows Based on Improved YOLO v3. Trans Chinese Soc Agric Mach 2020;51:250–60.
[24] Redmon J. Farhadi AJae-p. YOLOv3. An Incremental Improvement 2018.
[25] Yang S, Liu Y, Wang Z, Han Y, Wang Y, et al. Improved YOLO V4 model for face recognition of diary cow by fusing coordinate information 2021;37:129–35.
[26] Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: Optimal Speed and Accuracy of Object Detection. 2020.
[27] Deng J., Guo J., Ververas E., Kotsia I., Zafeiriou S., et al. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild.
[28] Yang S., Luo P., Loy C.C., Tang X. WIDER FACE: A Face Detection Benchmark. 2016.
[29] Schroff F., Kalenichenko D., Philbin J. FaceNet: A Unified Embedding for Face Recognition and Clustering. Proc. CVPR 2015.
[30] Huang G, Mattar M, Berg T, Learned-Miller E. Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments. Tech Rep 2008.
[31] Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. 2017.