

Original papers

Optimized BottleNet Transformer model with Graph Sampling and Counterfactual Attention for cow individual identification

Zhihao Xu, Yaqin Zhao^{*}, Zixuan Yin, Qiuping Yu

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Longpan Road, Nanjing, 210037, Jiangsu, China

ARTICLE INFO

Keywords:

Individual cow identification
Graph sampling
Counterfactual attention learning
Deep learning
Precision livestock

ABSTRACT

In modern dairy farms, accurate and reliable identification of each individual cow is of great significance for precision livestock farming. Individual cow identification is the basis for applications such as disease detection, automatic behaviour analysis, intelligent milking, and individual counting and is crucial for improving the welfare and breeding efficiency of dairy cows. Computer vision-based method is a low-cost, non-contact, automatic, and efficient way. To improve the accuracy and efficiency of cow recognition in different large-scale dairy farms, we proposed a BottleNet Transformer (BoTNet) model based on Graph Sampling and Counterfactual Attention Learning for cow surveillance videos. First, we replace the 3×3 spatial convolution with Multi-Head Attention in the final three bottleneck blocks of the ResNet. The BoT block module combines attention mechanisms and residual connection to enhance the global representation of cow images, which in turn better captures the features of the cow's back pattern region and ignores the influence of irrelevant information, such as the background of the dairy barn. Subsequently, counterfactual learning measures the quality of attention by comparing the difference between the generated output and the true label. The difference can be used to enhance the causal relationship between prediction results and cow feature attention, allowing the model to obtain more comprehensive cow appearance features. Finally, we added a Graph Sampling module before the feature extraction phase to produce small batches of samples for training. The GS sampler improves the learning efficiency while reducing the memory and computation consumption compared with the usual adopted PK sampling. We conducted comparison experiments on the public dataset Dataset1, and the experimental results reveal that the Rank-1, Rank-5, and mAP values of this study's method are 4%, 3.2%, and 5.3% higher than the optimal results, respectively, when compared with the existing state-of-the-art methods for animal individual recognition. In particular, we construct a challenging dataset by intercepting individual cow images from videos in the public dataset of farms. Experimental results indicate that the proposed method has better generalization performance.

1. Introduction

Dairy cows are one of the most common types of livestock in farming. Individual cow identification is of great significance for precision livestock farming, also known as integrated management systems. The identification is the basis for many applications, such as body condition scoring, disease detection, automatic behavior analysis, weighing, individual counting, intelligent milking, and dairy product traceability (Wang and Chen, 2023). In addition, the intelligent management of cow breeding, such as cow selection, vaccination, and calving, depends on individual cow identification (Kumar and Tiwari, 2016; Kumar et al., 2016).

The most common methods of automatically identifying cow individuals include marking ears or epidermis, wearing equipment such

as ear tags or e-collars (Wardrope, 1995), and radio frequency identification devices (Hong-da, 2012). However, these devices can cause adverse reactions. The insertion of ear tags may cause minor trauma, increasing the risk of infection. Furthermore, prolonged wearing of ear tags, leading to friction with the ear tissue, may induce discomfort in livestock, subsequently resulting in behavioral abnormalities (Edwards et al., 2001).

Video surveillance is a low-cost, noncontact, automatic, and efficient way to manage cows, and thus, in recent years, it has been widely used on farms (Xu et al., 2022; Yang et al., 2022; Beery et al., 2020). With the development of deep learning, many studies have attempted to use deep convolutional networks for individual cow recognition. Cow individuals vary in their facial, trunk, and tail head features that can

^{*} Corresponding author.

E-mail address: zhaoyaqin@njfu.edu.cn (Y. Zhao).

<https://doi.org/10.1016/j.compag.2024.108703>

Received 29 October 2023; Received in revised form 19 January 2024; Accepted 31 January 2024

Available online 8 February 2024

0168-1699/© 2024 Elsevier B.V. All rights reserved.

be used as identity information for cows. For instance, (Kumar et al., 2018) used a convolutional neural network with a deep belief network to extract cows' muzzle features for individual cow identification. When facing the issue of feature loss in cow face images due to changes in cow posture and different shooting angles, which causes a decline in the accuracy of individual cow recognition, Weng et al. (2022) used a two-branch convolutional neural network (CNN) to fuse the features of two cow face images taken at different angles. He et al. (2020) constructed a cow recognition model based on the improved YOLO v3 algorithm, specifically using the cow's back image as an input feature. To identify specific cows, Shen et al. (2020) employed the YOLO model to extract speckle features from the side view of cows, which in turn detects cow targets and then combines it with the improved AlexNet algorithm to recognize cows. Zhang et al. (2023) first located the torso region in the side-view walking image of a cow using the YOLOX model, then binarized the image with initial classification using the DeepOtsu model, and finally used the EfficientNet-B1 model to complete the final classification. Qiao et al. (2020) used Inception-V3 to extract the spatiotemporal features of the video sequence of a cow's backward glance and then combined them with a Bidirectional Long Short-Term Memory Network to identify the individual cow in the video.

Although the existing methods based on deep learning have achieved good recognition performance, these studies tend to view dairy barns as discriminative regions when identifying individual cows. Namely, these methods are suitable for dairy farms in the same context, and the recognition performance decreases significantly once the model is applied to other backgrounds. Furthermore, large dairy farms generate large-scale and diverse samples, but current methods cannot balance recognition accuracy and time consumption when processing them.

Bottleneck Transformer (BoTNet) is a network architecture that fuses CNN and Transformer (Srinivas et al., 2021). CNN is used to extract local features, whereas the Transformer block is used to extract global features. Counterfactual Attention Learning (CAL) (Rao et al., 2021; Pearl, 2022; VanderWeele, 2015) is introduced into the BoTNet network to adapt to the actual environment of large-scale farms. CAL provides a tool to evaluate the quality of attention and supervise the learning process by comparing the effects of the cow's back features and the background on the prediction results, which enhances the causal relationship between the prediction results and the attention to the local features of the cow's body, allowing the model to obtain more comprehensive appearance features. In addition, we use Graph Sampling (Liao and Shao, 2022) to build a nearest-neighbor relationship graph for each cow category to sample a small batch among the K most similar categories during the training period. Consequently, it is more conducive to providing informative and challenging instances for the discriminative learning of the individual cows, which will in turn improve the efficiency of models on large-scale datasets.

In summary, the innovations of this study are as follows. (1) We present a BoTNet model for individual cow recognition based on Graph Sampling and CAL, achieving a balance between computational consumption and recognition performance. The model can be applied to different large-scale dairy farms. (2) We also propose a solution to better capture the patterns on the cows' backs and minimize interference from irrelevant information, such as the background of the cow shed. Specifically, we make use of a BoTNet model to strengthen its global expression capability, combined with CAL to enhance the causal relationship between the prediction results and attention to the local features of the cow's body. Ultimately, our approach can extract more comprehensive appearance features from the cow. (3) To adapt to the large-scale dataset and diverse training samples of farms, we adopt Graph Sampling for small-batch sampling during the training period, which reduces memory and computation while improving efficiency compared with the PK sampling strategy usually adopted for deep learning.

Table 1

Statistical information for the DataSet1 dataset.

Datasets	Train		Test			
	Train IDs	Train images	Query IDs	Query images	Gallery IDs	Gallery images
FriesianCattle2017	88	568	49	101	71	238

Table 2

Statistical information for the DataSet2 dataset.

Datasets	Train		Test			
	Train IDs	Train images	Query IDs	Query images	Gallery IDs	Gallery images
Public Dataset	40	276	23	30	37	91

2. Experimental datasets

Since there are fewer studies on animal individual recognition, we adopt the dataset division criteria that are widely used in human recognition. We divide the dataset into a training set and a test set, dividing the test set into a gallery set and a query set. The query set refers to the dataset of images of cows with known identity numbers, and the gallery set refers to the dataset of images of cows with identity numbers that need to be determined by the model recognition.

We use two datasets to validate the model's performance. In DataSet1, the FriesianCattle2017 dataset from the University of Bristol <https://data.bris.ac.uk/data/dataset/2yizcfbkuv4352pzc32n54371r> was used for individual cow recognition research. The FriesianCattle2017 dataset consists of 907 top-view images of 84 individual cows taken in a milking parlor. To further evaluate the model's generalization ability and practical application, we also downloaded a public dataset: a video of cows taken on a dairy farm <https://doi.org/10.5281/zenodo.3981400>. We manually intercepted a video clip from a different dairy barn, constructing another dataset named DataSet2. The DataSet2 contains two different environments, day and night, totaling 40 cows and 407 images. Figs. 1 and 2 show the sample images from the two datasets. Tables 1 and 2 illustrate the dataset division.

3. Methods

Fig. 3 shows the architecture of the cow individual recognition network. First, we used the Graph Sampling (Liao and Shao, 2022) strategy to create a nearest-neighbor relationship graph for each cow category to sample small batches between the most similar K categories. Subsequently, we use the BoTNet network to extract the cow's back's local features and global contextual features, generating a multiscale feature map. Finally, CAL was employed to create a counterfactual intervention attention map. We could evaluate the quality of attention and improve the performance of recognition by measuring the difference between the original attention map generated by the BoTNet and the counterfactual intervention attention map.

3.1. Multiscale feature representation of cow's back using Botnet

As shown in Fig. 3, In the process of extracting the image features using the BoTNet network, the features are gradually extracted from low to high dimensions by gradually upscaling the structure through four layers, three Bottleneck layers, and one BoT Block. The output of the convolutional layers is then converted into a fixed-length feature vector through an average pooling operation. This feature vector is passed to a fully connected layer to generate the final feature map X.

Fig. 4 shows that the bottleneck structure in ResNet consists of three convolutional and BN layers. First, the number of input channels is reduced by using a 1×1 convolutional layer. Subsequently, a 3×3 convolutional layer is employed to increase the depth of the feature mapping. Finally, a 1×1 convolutional layer is utilized to recover the



Fig. 1. Sample of images from the dataset DataSet1.

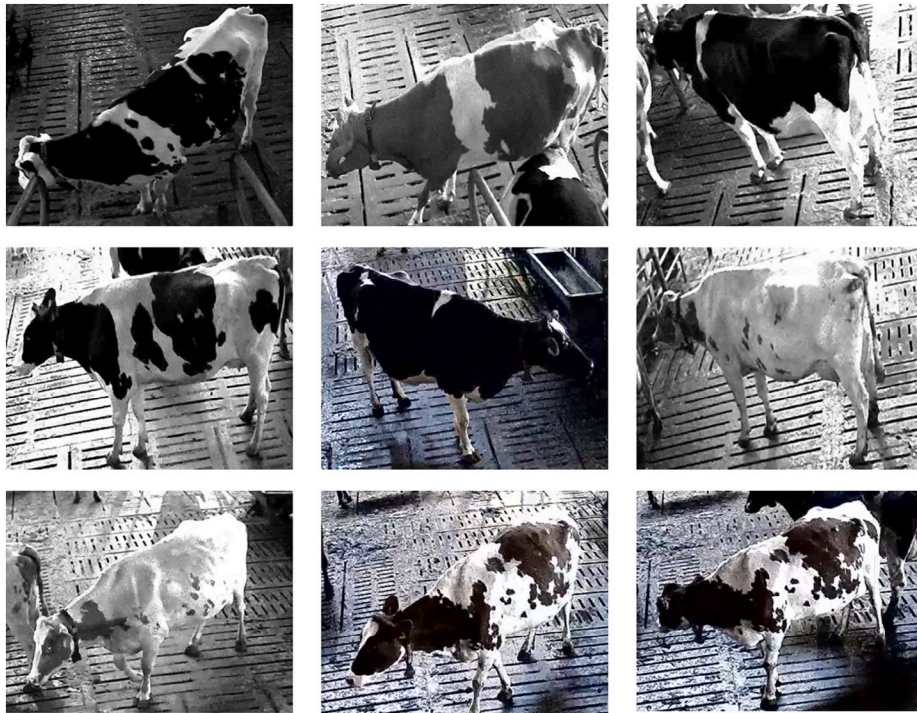


Fig. 2. Sample of images from the dataset DataSet2.

dimension of the feature map. The bottleneck structure introduces a residual structure if the input size does not match the output size. The residual structure reduces the size of the sampled input by a 1×1 convolutional layer in the downsample to establish the output size as the same as the input size.

The number of channels is first reduced by 1×1 convolution to map the features to a smaller feature space in the BoT block. Subsequently,

the BN layer normalizes the input and speeds up the convergence. Multi-Head Self-Attention (MHSA) is introduced to capture the dependencies between input features. MHSA can simultaneously focus on features at different spatial locations and learn long-distance dependencies that improve global representation. Once again, a BN layer is used to normalize the output, further enhancing the convergence of the network.

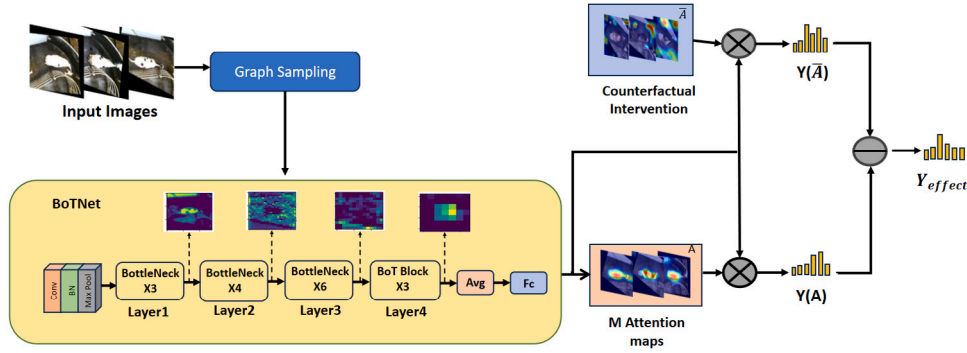


Fig. 3. Individual cow recognition network architecture. First, we use a Graph Sampling strategy to generate small batches of samples for training. Subsequently, we extract multiscale feature maps using the BoTNet network. Finally, we apply counterfactuals to intervene in the original attention graph. We analyze the effect of learning visual attention through the difference between the original and counterfactual classification results and maximize this difference during training to improve the model's performance.

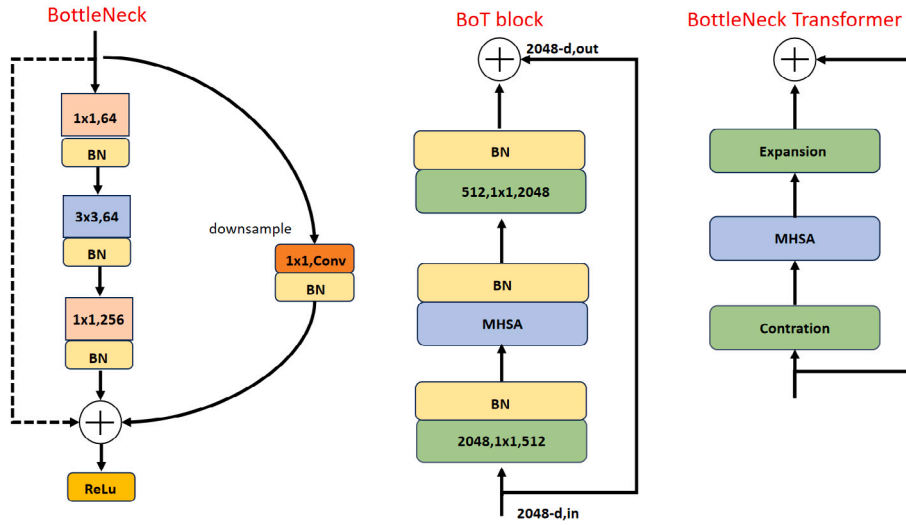


Fig. 4. Left: BottleNeck block; middle: BoT block; right: BottleNeck Transformer module. A ResNet bottleneck block with an MHSA layer is similar to a Transformer structure and can be seen as having a similar function.

Finally, another 1×1 convolution recovered the number of channels, and a BN layer normalized the features. Since ResNet introduced MHSA, feature interaction capability similar to BoTNet is achieved through self-attention. Consequently, BoT blocks can be considered BoT Nets with similar functions.

Extracting the texture features of cows is key to individual recognition. Traditional CNNs can only extract local features, but local features alone may not accurately capture the characteristics and contextual information of the cow's entire body texture. Therefore, global features also play an important role in discriminating individual cows. For the task of recognizing the cow's back pattern, we extract local features at different locations and scales on the cow's back through a CNN. The Transformer block performs information interaction and feature extraction over the whole image through a self-attention mechanism to capture global features. Specifically, the block consists of a ResNet bottleneck with Multi-Head Attention layers, which effectively captures the semantic correlation between features and enhances the ability of global feature representation by simultaneously considering different parts of the input features and learning the relationship between them. Finally, the BoTNet network fuses local features with global features.

3.2. Enhancing cow characteristics: CAL

The CAL framework is an attention optimization method for fine-grained visual classification and reidentification. CAL can help optimize the quality of attention in the tasks of individual cows. Therefore, CAL

adjusts its attention more to these key feature regions after learning the features of individual cows through the model.

When inputting a picture of a cow, the multiscale feature map X learned by BoTNet is subjected to a dot-multiplication operation with the correct attention map A to obtain another new feature map W_1 in order to improve the model's feature representation in the region of interest. By assigning a specific value to a variable through a counterfactual intervention (e.g., $do(A = \bar{A})$), assigning variable A to a fixed value \bar{A} and eliminating all preconditions for that variable (e.g., $X \rightarrow A$, variable A is affected by precondition X , and by intervening so that variable A is no longer affected by precondition X), under the condition that the feature map is unchanged, we produce an attention map \bar{A} of the counterfactual intervention for individual cow identification. We dot-multiply the attention map of the counterfactual intervention with the original feature map X to obtain another new feature map W_2 , after which we predict the outcomes $Y(A)$ and $Y(\bar{A})$, respectively, and by comparing the correct prediction of attention (i.e., the cow texture region) with the incorrect prediction of attention (i.e., the dairy farm background), we can assess the quality and effectiveness of the acquired attention. The formula is calculated as follows:

$$Y_{effect} = E_{\bar{A} \sim \gamma} [Y(A) - Y(\bar{A})] \quad (1)$$

Y_{effect} refers to the effect of the disparity in accurate and inaccurate attention on the anticipated outcome. Additionally, γ represents the distribution of CAL.

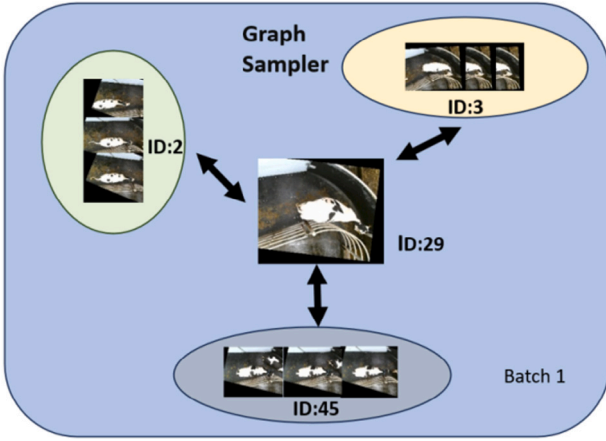


Fig. 5. Flow of Graph Sampling. There are 88 cows in total. We select a particular individual cow as anchor c (e.g., ID: 29) and retrieve all its relation classes among the 88 cows (e.g., ID: 2, ID: 3, ID: 45).

Attention is optimized by formulating the objective function to improve the prediction results as follows (Rao et al., 2021):

$$L = L_{ce}(Y_{effect}, \gamma) + L_{others} \quad (2)$$

L_{ce} represents the cross-entropy loss, γ represents the classification labels, and L_{others} represents the original objective function.

3.3. Network training based on graph sampling

One common method for random, small-batch sampling is called PK sampling during deep learning training. The PK sampler randomly selected P categories for each small batch and then randomly sampled K samples within each category, which can ensure diversity and balance in the sample selection process. However, complete random sampling may not provide sufficient information for discriminative learning, whereas Graph Sampling preserves the relevant information between samples, which is beneficial for extracting more generalizable feature representations. Graph Sampling also reduces computational loss and increases the speed of training and inference for the individual cow identification model, which helps the model be applied to different dairy farms.

The Graph Sampling strategy used in this study is shown in Fig. 5. First, we randomly select a graph of a particular cow to construct a subdataset. Subsequently, for a particular cow individual, we retrieve $P-1$ similar individuals to construct a relationship graph G (P is the number of categories sampled under each small batch). Specifically, feature embeddings are extracted from the subdataset $X \in R^{C \times d}$, where C is all categories of cows and d is the feature dimension. Subsequently, QAConv obtains all distance matrices by means of calculating pair-wise distances for all samples. $P-1$ similar (P is the category of individual cows) is retrieved to construct $G(V, E)$ for a certain individual cow, where $V = \{c|c = 1, 2, \dots, C\}$ denotes nodes constituted by individual cows. $E = \{(c1, c2)|c2 \in N(c1)\}$ stands for edges and selects a specific class of individual cows as an anchor point c , retrieving all its relation classes in G and obtaining the set A ($|A| = P$). The total number of set A is P . K instances are randomly sampled in each class in A to generate samples for training.

4. Experiments

4.1. Experimental details

As shown in Table 3, we used Python 3.6 to develop the algorithms for this experiment and implemented them on the PyTorch 1.71 framework. The hardware setup consisted of an NVIDIA RTX 3060 GPU,

Table 3

Experimental environment parameters.

Parameter setting	Version
Operating system	Windows11
CPU	Intel(R) Core(TM) i7-12700H
GPU	NVIDIA RTX 3060 GPU
RAM	DDR4
PyTorch	1.7.1

an Intel(R) Core(TM) i7-12700H CPU, and the Windows 11 operating system for training and testing.

We adopted the standard BoTNet50 as the backbone network, and the input image size is 384×192 . Several data enhancement methods are applied to augment the number of training samples, such as random cropping, masking, horizontal flipping, and mirroring. All experiments are conducted with the same hyper-parameters, including 10 image padding, 16 batch sizes, and 0.9 momentum. The training period is 60 epochs, with an initial learning rate of 0.0001. During training, the network utilizes the Adam optimizer and triplet loss function. Cumulative Matching Characteristics (CMC) and Mean Average Precision metrics are used for evaluating the experimental results, where we used Rank-1 and Rank-5 to evaluate the CMC, denoting the first hit accuracy and the first five image matching accuracy.

4.2. Cow identity matching and inquiry

We demonstrated the visualization of the model's predicted results (Rank-5). As shown in Fig. 6, the left image shows an image of a cow with the identity number to be queried in the query, and the right image shows an image of a cow with an unknown identity number in the gallery. We aim to find cows in the gallery belonging to the same identity number as the query by the trained model and verify its accuracy. In the case of a given query image, Rank-1 denotes the most similar image returned by the model. According to the similarity ordering, Rank-5 denotes the top 5 most similar images. In summary, Rank-1 and Rank-5 are essential metrics for evaluating model performance.

We visualize the results of the first five matches for individual cow identification, where green boxes indicate correct predictions and red boxes indicate incorrect predictions. The results reveal that only cow (c) is incorrect in Rank-5 image matching, suggesting that our method performs well on Rank-1 and Rank-5 and can accurately identify individual cows.

4.3. Comparison with the PK sampler

As shown in Table 4, to better verify the effectiveness of the GS sampler in the network, we compare two minibatch sampling methods, including PK and GS. The experimental results demonstrate that the proposed GS sampler is better than the PK sampler for different backbone networks. Specifically, for ResNet50, the GS sampler is 2.2%, 1%, and 1.9% higher on Rank-1, Rank-5, and mAP, respectively. For BoTNet50, the GS sampler is 2.2%, 1.1%, and 1.2% higher on Rank-1, Rank-5, and mAP, respectively. The GS sampler can find classes with high similarity as difficult examples to enhance learning ability. It is beneficial for discriminative models to extract features by comparing the differences and commonalities between these confusing examples. Therefore, the GS sampler is more efficient than the PK sampler in terms of its ability to learn the features of cows.

4.4. Comparison with advanced attention mechanisms

To validate the effect of the counterfactual intervention attention mechanism in the cow individual recognition network proposed in this study, we replace the counterfactual intervention attention mechanism with several popular attention mechanisms. To ensure the fairness of

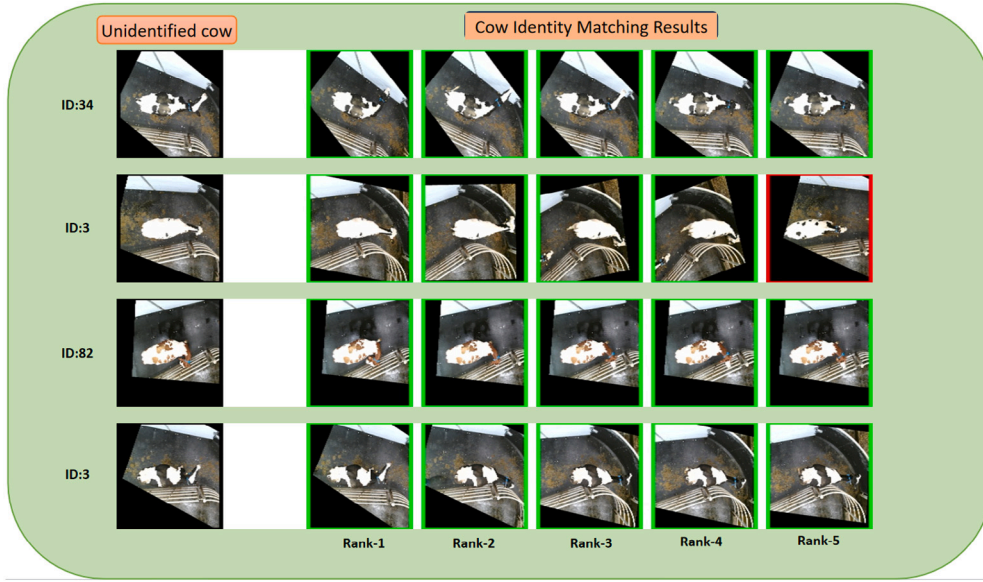


Fig. 6. The picture shows the visualization of Rank-5 accuracy. Left: picture of an unidentified cow. Right: results of the first five matches for the cow's identity. The green box represents a correct prediction, and the red box represents an incorrect prediction.

Table 4
Comparison of different batch samplers.

Method	Datasets			
	Model	R1	R5	mAP
PK	ResNet 50	93.5%	96.8%	94.3%
	BoTNet50	94.6%	97.8%	95.8%
GS	ResNet 50	95.7%	97.8%	96.2%
	BoTNet50	96.8%	98.9%	97.0%

Table 5
Comparison with advanced attention mechanisms.

Method	Datasets		
	R1	R5	mAP
GS+CBAM	94.1%	97.8%	94.9%
GS+CA	95.7%	96.8%	93.0%
GS+GAM	94.4%	97.8%	94.3%
GS+NAM	95.7%	97.8%	94.7%
QAConv	93.5%	96.8%	93.8%
GS+CAL(ours)	96.8%	98.9%	97.0%

the experiment, we followed the same hyper-parameter and network structure settings for several individual cow identification methods. As shown in Table 2, we conducted comparative experiments on five mainstream attention mechanisms, including QAConv + SE (Liao, 2020), NAM (Liu and Shao, 2021), GAM (Liu et al., 2021), CA (Hou et al., 2021), and CBAM (Woo et al., 2018). We can see from Table 5 that the attention learned by these methods is not always optimal, and in particular, CBAM and GAM do not improve model stability significantly. Our method achieves the same Rank-1 as GAM and NAM, which is 95.7%. Although Rank-1 is a vital evaluation metric, mAP more fully reflects the quality of the model. Our method outperforms GAM and NAM in mAP by 2.1% and 2.7%, respectively. Unlike other attention mechanisms, we compare correct and incorrect attention through the CAL framework to infer a more accurate attention distribution. Therefore, our method can optimize the learned attention to obtain optimal results on mAP.

We observe that Fig. 7 illustrates the dispersion of the heat map of attention. The attention heatmap of QAConv in Fig. 7(b) cannot focus well on the regions that distinguish the cows from the background, suggesting poor dispersion. However, as shown in Fig. 7(c), the CAL-optimized attention method exhibits better focus on regions of the cow

texture. Nevertheless, it still incorrectly focuses on certain background regions. As shown in Fig. 7(d), using BoTNet with Multi-Head Attention and combining it with CAL can further focus attention on the texture region of the cow.

4.5. Validation of the BoTNet structure

In addition, we focus on the training efficiency of the proposed method. As shown in Fig. 8, the BoTNet50 network improves the Rank-1 metric by 1.1% and the mAP metric by 0.8% compared with the ResNet50. In particular, BoTNet50 outperforms not only ResNet50 but also ResNet152, with a 0.7% higher average accuracy mAP and better model stability. The BoTNet50 network employs Multi-Head Attention instead of traditional spatial convolution, which automatically focuses on the dependencies of different regions of the cow's back and improves the feature representation of the model. Compared with ResNet50, BoTNet50 reduces the number of parameters by reducing the number of output channels per convolutional layer, generating a dimensionality reduction of the output feature map per convolutional layer. Although BoTNet50 has fewer parameters and is more lightweight, its accuracy is comparable to that of deep networks.

Conversely, traditional PK samplers need to calculate and store the pair-wise distance between many samples (the similarity between samples). The GS sampler uses nearest-neighbor sampling and thus only needs to compute the pair-wise distance of some of the samples, reducing the training time eventually. As shown in Figs. 9 and 10, our method deals with each small batch of data in an average time of 19.5 s and fewer parameters, which is better than others.

4.6. Comparison with state-of-the-art technology

4.6.1. Comparison with existing methods of individual animal identification

We compared the method proposed with existing methods for animal individual identification, named tigerReID (Yu et al., 2019), Part-Pose ReID (Liu et al., 2019), and PrimNet (Deb et al., 2018). The experimental results are shown in Table 6, where the Rank-1, Rank-5, and mAP values of the presented method are 4.0%, 3.2%, and 5.3% higher than the optimal results, respectively, of the three compared methods. We observe that our method performs better than the other three methods. TigerReID is a method that combines global and local features through deep convolutional networks. However, it needs to

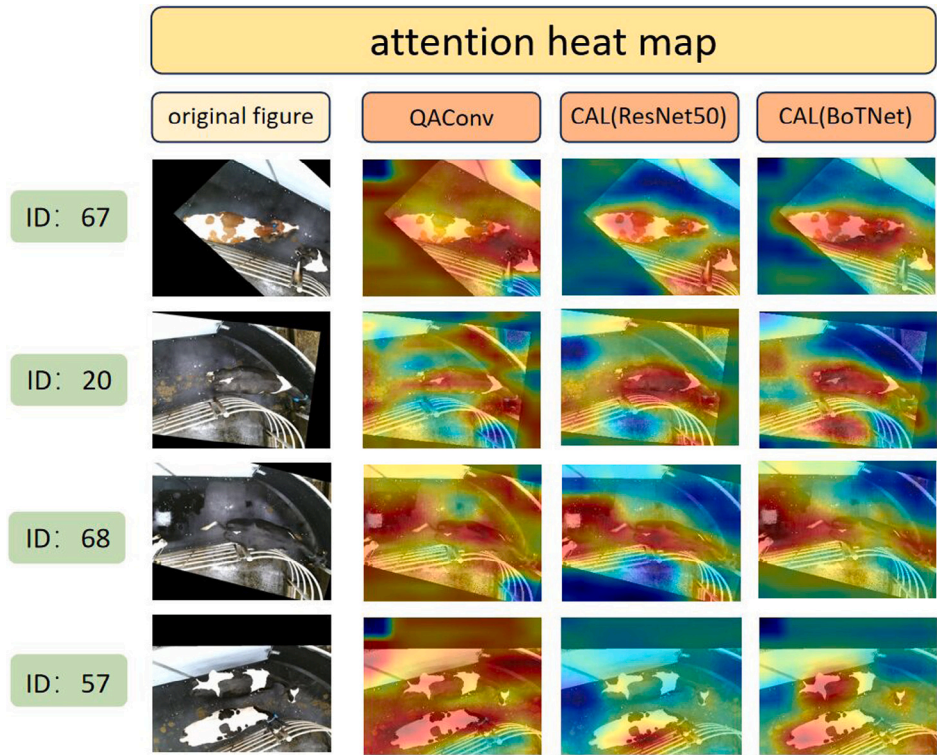


Fig. 7. Attention heatmap of dairy cows: (a) original image; (b) QAConv attention heatmap; (c) CAL (ResNet50) attention heatmap; and (d) CAL (BoTNet) attention heatmap (ours). The pictures show the distraction of the heat map of attention presented by the different methods.

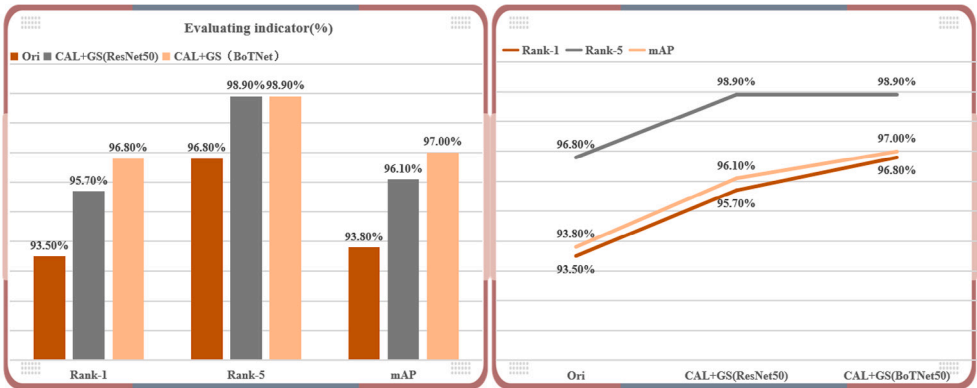


Fig. 8. Results of comparison with other models. The pictures show the recognition accuracy of the different methods. Rank-1 denotes the most similar image returned by the model. According to the similarity ordering, Rank-5 denotes the top 5 most similar images. MAP denotes the mean average precision of the model.

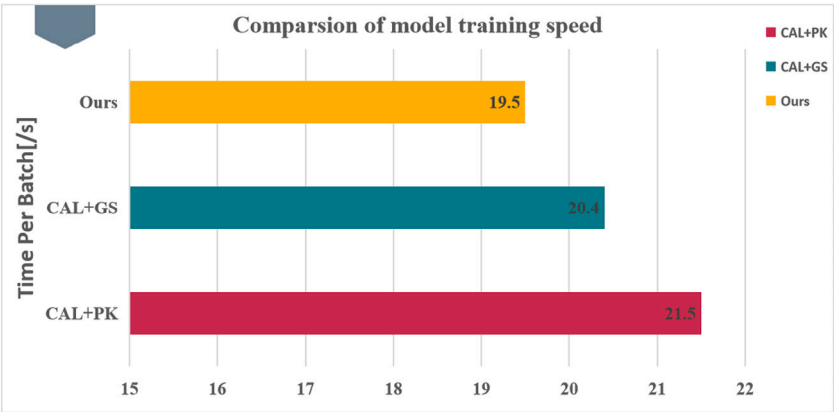


Fig. 9. Comparison of model training speeds. The image shows the comparison between different samplers for each small batch of training speeds.

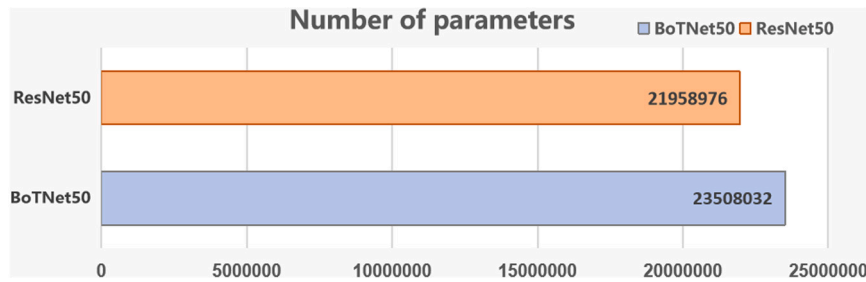


Fig. 10. Comparison of model parameters. The picture shows the number of parameters of the two different backbone models.

Table 6

Comparison with existing methods of individual animal identification.

Method	Datasets		
	R1	R5	mAP
tigerReID	90.1%	94.3%	87.5%
Part-Pose ReID	92.8%	95.7%	91.7%
PrimNet	79.8%	–	–
CAL+GS(Resnet50)	95.7%	98.9%	96.1%
CAL+GS(Resnet152)	96.8%	98.9%	96.3%
CAL+GS(BoTNet50)	96.8%	98.9%	97.0%

Table 7

Comparison with different person re-identification methods.

Method	Datasets		
	R1	R5	mAP
BDB	91.4%	96.8%	89.7%
Top-DB-Net	92.5%	97.8%	92.2%
OSNet	94.2%	96.8%	94.5%
NFormer	95.7%	97.8%	95.8%
MSINet	91.4%	96.8%	91.2%
OSNet-AIN	95.7%	98.9%	96.1%
CAL+GS(Resnet50)	95.7%	98.9%	96.1%
CAL+GS(Resnet152)	96.8%	98.9%	96.3%
CAL+GS(BoTNet50)	96.8%	98.9%	97.0%

pay attention to the importance of the attention mechanism for selecting feature regions. Instead, we focus more on the regions that are more helpful in distinguishing individuals through the attention mechanism. The cows in our dataset were in a lying position, whereas Part-Pose ReID required the animals to be in a standard standing position to obtain head, back, and leg features. Owing to the influence of factors such as light and occlusion in a dataset, it is difficult for our datasets to provide clear facial images for PrimNet, which reduces the accuracy and stability of recognition.

4.6.2. Comparison with advanced human individual identification methods

Although there are few studies on individual animal recognition, we can draw inspiration from person reidentification studies. Our study borrowed the structure of BoTNet, a network for person reidentification, to construct a model for individual cow recognition. We compare our method with the state-of-the-art (SOTA) methods for person reidentification to Zhou et al. (2021) verify the model's performance, such as MSINet (Gu et al., 2023), OSNet (Zhou et al., 2019), BDB (Dai et al., 2019), Top-DB-Net (Quispe and Pedrini, 2021), NFormer (Wang et al., 2022), and OSNet-AIN (Zhou et al., 2021). The experimental results are shown in Table 7, which shows that this study outperforms the optimal result method by 1.1% in Rank-1, and the mAP value is 0.9% higher than the optimal result. Owing to the efficient learning efficiency of the GS sampler, the BoTNet model combines the advantages of CNN and Transformer to effectively extract local features and improve global expression, and finally, CAL optimizes the results of attention learning. Our model can still provide satisfactory predictions even when facing cows with similar colors.

5. Discussion

Accurate detection and localization of individual cows are crucial prerequisites for identity recognition, and individual cow recognition technology can be applied not only for livestock tracking and counting but also for animal behavior analysis to monitor animal welfare. Inspired by the SOTA person-identification architecture, QAConv, this study combines CAL and Graph Sampling to present the model's performance in identifying individual cows in real-world scenarios. Notably, the environment for each cow is unique. As shown in Table 8, compared with the existing methods for individual livestock recognition, the proposed approach exhibits the following advantages.

In the context of individual cow identification, conventional approaches typically involve full-sample training, resulting in a substantial computational burden and being applicable only to particular backgrounds. Conversely, small-batch sampling methods introduce a degree of unpredictability during the training process. This unpredictability serves to mitigate the model's tendency to overfit specific samples and enhance its robustness. Nonetheless, the randomness of small-batch sampling may also result in inadequate information acquisition. The GS sampler effectively resolves this concern by establishing a nearest-neighbor graph to identify similar samples for training purposes.

5.1. More suitable for large-scale farms

Current methods for individual cow recognition input a large number of samples simultaneously during training, which increases the computational cost despite achieving high recognition accuracy. However, the manner in which to balance computational consumption with recognition accuracy is relevant for processing cow data from large farms. Small-batch sampling randomly divides the large-scale dataset into different small batches and processes a portion of these samples at a time, which significantly reduces computational consumption and improves training efficiency. The study adopts a small-batch metric learning strategy based on Graph Sampling to train the model. Different from common PK sampling (Liao and Shao, 2022), the Graph Sampling method constructs the nearest-neighbor relationship graph by calculating the pair-wise distances of all the samples. Subsequently, challenging samples with high similarity are selected for training in small batches, which enables the model to focus on mining the commonalities of the same individual cows, reduces noise interference, and improves the model's generalization ability. Furthermore, large-scale datasets may not be feasible to load entirely into memory. However, small-batch sampling only requires processing the current batch of data, which reduces the model's memory requirements.

5.2. Applied to cowsheds with different backgrounds

To the best of our understanding, the majority of livestock individual recognition techniques are developed and evaluated for specific contexts in which the background of cow images remains consistent.

Table 8
Comparison of different batch samplers.

Methods	Data sampling methods	Availability of attention mechanisms	Generalization capability In different backgrounds	Positioning and video identification	Data sampling methods
Traditional methods	Full-sample training	✗	Applicable to specific contexts	✗	High
Part-Pose ReID	Full-sample training	✗	Inferior	✓	High
QAConv	PK sampling	✗	Better	✗	Low
OSNet-AIN	Full-sample training	✓	Better	✗	High
Ours	GS sampling	✓	Better	✓	Low

Table 9
Comparison with other advanced identification methods.

Method	Dataset1			Dataset2		
	R1	R5	mAP	R1	R5	mAP
OSNet-AIN	95.7%	98.9%	96.1%	86.3%	95.8%	85.9%
Part-Pose ReID	92.8%	95.7%	91.7%	82.7%	92.6%	81.7%
QAConv	93.5%	96.8%	93.8%	83.1%	93.1%	84.7%
Ours	96.8%	98.9%	97.0%	87.6%	94.7%	86.1%

However, when these models are applied to barns with varying backgrounds, achieving high recognition accuracy becomes challenging. Some human individual recognition methods seek to improve the performance of detailed recognition tasks by incorporating attention mechanisms. These mechanisms enable the model to prioritize regions in the image that are beneficial for recognition. Nevertheless, these methods often adjust attention mechanisms based on the final recognition outcomes. In our work, we utilize a counterfactual learning approach, which involves comparing the disparities between attention maps with added counterfactual perturbations and original attention maps during the training phase. This strategy aims to assess the feature extraction performance of the cow trunk region in the original attention map. Subsequently, adjustments are made to focus attention on specific regions, thereby reducing interference from background areas in the barn. Consequently, this methodology enhances the resilience and adaptability of the individual cow recognition task. Our dataset, as depicted in Figs. 1 and 2, is sourced from diverse scenes. Cross-dataset testing employs different datasets for training and testing to assess the model's performance. This approach enables us to evaluate the models' generalization ability across diverse environments. To ensure the fairness of the experiments, our trials will be trained on 907 cow images from DataSet1. Subsequently, cross-dataset testing will be performed on 397 cow images in DataSet2 recorded from different cowsheds on the farm. Table 9 shows that our approach achieves favorable accuracy across different backgrounds.

5.3. Applications

The process of realizing individual cow recognition in videos involves two key steps: object detection and identity recognition. Initially, we utilize the YOLOv5 object detection framework to quickly locate cows in video frames and obtain their positional information. Once YOLOv5 successfully detects and locates the cows, we apply our method to extract the texture features of the cows. Finally, the extracted cow texture features are compared with the known cow texture features in the database. The identity of individual cows is ultimately recognized by comparing the similarity between samples (see Fig. 11).

To assess the practical applicability of the algorithmic model, we conducted tests using cow videos from the publicly available DataSet2.

These videos have a resolution of $1,920 \times 1,080$ pixels and a frame rate of 25 fps. The method proposed in this study successfully retrieved and identified individual cows from the videos when provided with one or more images of cow identities. A sample of this process is illustrated in Fig. 11. This capability enables the analysis of daily cow behavior and health monitoring. Additionally, individual cow identification facilitates intelligent milking and enhances the traceability of dairy products (see Fig. 11).

5.4. Limitations

Although the individual cow identification model proposed in this study can accurately recognize the identities of cows in the videos, in practical applications, the trunk features of cows may be obscured by other cows, or specific poses such as curling may result in the inability to capture the entire part of the trunk. Furthermore, the trunk features of cows may undergo visual changes due to soiling or growth. In such scenarios, the model's performance may degrade. Fig. 12 provides samples of identification results in these situations.

6. Conclusion

In the current study, we propose a BoTNet model based on Graph Sampling and CAL optimization for individual cow recognition. Specifically, we first use the GS sampler to select valuable small batches of samples, which can reduce training time and increase the challenge. Subsequently, we use the BoTNet model for feature extraction. The BoTNet model improves the global feature expression of the image through BoT blocks to better recognize individual cows. Finally, we use CAL to optimize the quality of attention learning. In the task of individual cow recognition, our method performs better by analyzing the experimental results, achieving a balance between computational consumption and recognition performance, and can apply to other large-scale dairy farms.

Although our method has improved accuracy significantly, there are still limits to it when faced with cows with little difference in mottling in the image and in the case of multiple cow occlusions. We have decided to address these issues in the next step. According to research, Transformer is a neural network model suitable for global feature extraction, whereas CNN has the advantage of extracting local features. In summary, we consider processing the input in parallel using CNN and Transformer and then fuse the feature obtained. Finally, we further process the new features obtained from fusion.

In addition, we expect to be able to learn new feature representations based on the existing model structure rather than training from scratch. In future studies, to address the issue of variation in cow trunk features, we will attempt to introduce an incremental recognition model to improve its adaptability.



Fig. 11. Sample image illustrating the outcomes of locating and querying an individual cow. Upper left corner: identification number. Lower left corner: time information.



Fig. 12. Scenario with poor recognition results.

CRediT authorship contribution statement

Zhihao Xu: Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. **Yaqin Zhao:** Methodology, Writing – review & editing. **Zixuan Yin:** Data curation. **Qiuping Yu:** Software.

Declaration of competing interest

All authors disclosed no relevant relationships.

Data availability

Data will be made available on request.

Acknowledgments

Supported by National Natural Science Foundation of China (32371583).

References

- Beery, S., Wu, G., Rathod, V., Votel, R., Huang, J., 2020. Context r-cnn: Long term temporal context for per-camera object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13075–13085.
- Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P., 2019. Batch dropblock network for person re-identification and beyond. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3691–3701.
- Deb, D., Wiper, S., Gong, S., Shi, Y., Tymoszek, C., Fletcher, A., Jain, A.K., 2018. Face recognition: Primates in the wild. In: *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems*. BTAS, IEEE, pp. 1–10.
- Edwards, D.S., Johnston, A.M., Pfeiffer, D.U., 2001. A comparison of commonly used ear tags on the ear damage of sheep. *Animal Welf.* 10 (2), 141–151.
- Gu, J., Wang, K., Luo, H., Chen, C., Jiang, W., Fang, Y., Zhang, S., You, Y., Zhao, J., 2023. MSINet: Twins contrastive search of multi-scale interaction for object ReID. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19243–19253.
- He, D., Liu, J., Xiong, H., Lu, Z., 2020. Individual identification of dairy cows based on improved YOLO v3. *Trans. Chin. Soc. Agric. Mach.* 51 (4), 250–260.
- Hong-da, W., 2012. Application of radio frequency identification (RFID) in dairy information management. *J. Northeast Agric. Univ. (Engl. Ed.)* 19 (1), 78–81.
- Hou, Q., Zhou, D., Feng, J., 2021. Coordinate attention for efficient mobile network design. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 13713–13722.
- Kumar, S., Pandey, A., Satwik, K.S.R., Kumar, S., Singh, S.K., Singh, A.K., Mohan, A., 2018. Deep learning framework for recognition of cattle using muzzle point image pattern. *Measurement* 116, 1–17.
- Kumar, S., Singh, S.K., Dutta, T., Gupta, H.P., 2016. Poster: A real-time cattle recognition system using wireless multimedia networks. In: *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion*. pp. 48–48.
- Kumar, S., Tiwari, S., 2016. Face recognition of cattle: Can it be done? *Proc. Natl. Academy Sci. India Sect. A: Phys. Sci.* 86.
- Liao, S., 2020. Interpretable and generalizable person re-identification with query-adaptive convolution and temporal lifting. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI* 16. Springer, pp. 456–474.
- Liao, S., Shao, L., 2022. Graph sampling based deep metric learning for generalizable person re-identification. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7359–7368.
- Liu, Y., Shao, Z., 2021. NAM: Normalization-based attention module. *arXiv preprint arXiv:2111.12419*.
- Liu, Y., Shao, Z., Hoffmann, N., 2021. Global attention mechanism: Retain information to enhance channel-spatial interactions. *arXiv preprint arXiv:2112.05561*.
- Liu, C., Zhang, R., Guo, L., 2019. Part-pose guided amur tiger re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- Pearl, J., 2022. Direct and indirect effects. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. Oxford University Press, pp. 373–392.
- Qiao, Y., Su, D., Kong, H., Sukkari, S., Lomax, S., Clark, C., 2020. BiLSTM-based individual cattle identification for automated precision livestock farming. In: *2020 IEEE 16th International Conference on Automation Science and Engineering*. CASE, IEEE, pp. 967–972.
- Quispe, R., Pedrini, H., 2021. Top-db-net: Top dropblock for activation enhancement in person re-identification. In: *2020 25th International Conference on Pattern Recognition*. ICPR, IEEE, pp. 2980–2987.
- Rao, Y., Chen, G., Lu, J., Zhou, J., 2021. Counterfactual attention learning for fine-grained visual categorization and re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1025–1034.
- Shen, W., Hu, H., Dai, B., Wei, X., Sun, J., Jiang, L., Sun, Y., 2020. Individual identification of dairy cows based on convolutional neural networks. *Multimedia Tools Appl.* 79, 14711–14724.
- Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A., 2021. Bottleneck transformers for visual recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16519–16529.
- VanderWeele, T., 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press.
- Wang, J., Chen, H., 2023. Identification of oestrus cows based on vocalisation characteristics and machine learning technique using a dual-channel-equipped acoustic tag. *animal* 17 (6), 100811.
- Wang, H., Shen, J., Liu, Y., Gao, Y., Gavves, E., 2022. Nformer: Robust person re-identification with neighbor transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7297–7307.
- Wardrope, D., 1995. Problems with the use of ear tags in cattle. *Veterin. Rec.* 137 (26), 675.
- Weng, Z., Meng, F., Liu, S., Zhang, Y., Zheng, Z., Gong, C., 2022. Cattle face recognition based on a two-branch convolutional neural network. *Comput. Electron. Agric.* 196, 106871.
- Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision*. ECCV, pp. 3–19.
- Xu, Y., Liu, J., Wan, Z., Zhang, D., Jiang, D., 2022. Rotor fault diagnosis using domain-adversarial neural network with time-frequency analysis. *Machines* 10 (8), 610.
- Yang, F., Jiang, Y., Xu, Y., 2022. Design of bird sound recognition model based on lightweight. *IEEE Access* 10, 85189–85198.
- Yu, J., Su, H., Liu, J., Yang, Z., Zhang, Z., Zhu, Y., Yang, L., Jiao, B., 2019. A strong baseline for tiger re-id and its bag of tricks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- Zhang, R., Ji, J., Zhao, K., Wang, J., Zhang, M., Wang, M., 2023. A cascaded individual cow identification method based on DeepOtsu and EfficientNet. *Agriculture* 13 (2), 279.
- Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2019. Omni-scale feature learning for person re-identification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3702–3712.
- Zhou, K., Yang, Y., Cavallaro, A., Xiang, T., 2021. Learning generalisable omni-scale representations for person re-identification. *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (9), 5056–5069.