frontiers
in Animal Science

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fanim.2021.759147

Automated Individual Cattle Identification Using Video Data: A Unified Deep Learning Architecture Approach

Yongliang Qiao [1], Cameron Clark [2], Sabrina Lomax [2], He Kong [1], Daobilige Su [3] and Salah Sukkarieh [1]*

[1] Australian Centre for Field Robotics, Faculty of Engineering, The University of Sydney, Sydney, NSW, Australia, [2] Livestock Production and Welfare Group (www.livestockproductionandwelfare.com), School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, Sydney, NSW, Australia, [3] College of Engineering, China Agricultural University, Beijing, China

OPEN ACCESS

Edited by:
Suresh Neethirajan,
Wageningen University and Research,
Netherlands

Reviewed by:
Pablo Fonseca,
University of Guelph, Canada
Mehmet Gültas,
South Westphalia University of
Applied Sciences, Germany

****Correspondence:***
Yongliang Qiao
yongliang.qiao@sydney.edu.au

Specialty section:
This article was submitted to
Precision Livestock Farming,
a section of the journal
Frontiers in Animal Science

Received: *15 August 2021*
Accepted: *27 October 2021*
Published: *23 November 2021*

Citation:
Qiao Y, Clark C, Lomax S, Kong H,
Su D and Sukkarieh S (2021)
Automated Individual Cattle
Identification Using Video Data: A
Unified Deep Learning Architecture
Approach.
Front. Anim. Sci. 2:759147.
doi: 10.3389/fanim.2021.759147

Individual cattle identification is a prerequisite and foundation for precision livestock farming. Existing methods for cattle identification require radio frequency or visual ear tags, all of which are prone to loss or damage. Here, we propose and implement a new unified deep learning approach to cattle identification using video analysis. The proposed deep learning framework is composed of a Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) with a self-attention mechanism. More specifically, the Inception-V3 CNN was used to extract features from a cattle video dataset taken in a feedlot with rear-view. Extracted features were then fed to a BiLSTM layer to capture spatio-temporal information. Then, self-attention was employed to provide a different focus on the features captured by BiLSTM for the final step of cattle identification. We used a total of 363 rear-view videos from 50 cattle at three different times with an interval of 1 month between data collection periods. The proposed method achieved 93.3% identification accuracy using a 30-frame video length, which outperformed current state-of-the-art methods (Inception-V3, MLP, SimpleRNN, LSTM, and BiLSTM). Furthermore, two different attention schemes, namely, additive and multiplicative attention mechanisms were compared. Our results show that the additive attention mechanism achieved 93.3% accuracy and 91.0% recall, greater than multiplicative attention mechanism with 90.7% accuracy and 87.0% recall. Video length also impacted accuracy, with video sequence length up to 30-frames enhancing identification performance. Overall, our approach can capture key spatio-temporal features to improve cattle identification accuracy, enabling automated cattle identification for precision livestock farming.

Keywords: cattle identification, deep learning, BiLSTM, self-attention, precision livestock farming

1. INTRODUCTION

Cattle identification is the process of accurately recognizing individuals via a unique identifier or biometric feature(s) (Berckmans, 2014). In precision livestock management, individual cattle identification is a prerequisite for automated analysis of animal activities and productivity (Dawkins, 2021). Classical cattle identification methods typically adopt on-animal sensors such as ear-tags, collars, and radio frequency identification modules, which incur costs and may also burden cattle (Andrew et al., 2016). In addition, these tags or sensors are prone to loss or damage in harsh outdoor environments (Rotter, 2008). Therefore, a more robust cattle identification system of high accuracy is desirable.

Deep learning networks with automatic feature extraction and powerful image representation capability have been widely used in the field of object detection, visual recognition and image segmentation (Qiao et al., 2019a,c, 2021). As a result, there has been recent interest in the use of deep learning for cattle feature extraction and identification of individual animals (Kumar et al., 2018; Qiao et al., 2020). In existing approaches, deep learning models such as Convolutional Neural Networks (CNN) are utilized to extract high-dimensional visual features in a spatial domain from images, with these extracted features then being used to identify animals through a classifier layer. For example, de Lima Weber et al. (2020) used CNN for Pantaneira cattle breed recognition and achieved 99% accuracy in DenseNet-201, Resnet50, and Inception-Resnet-V. Andrew et al. (2017) used R-CNN deep neural network to determine coat characteristics for single frame-based individual cattle identification. Moreover, Shen et al. (2019a) extracted cow trunk using the YOLO detection model and then used a fine-tuned AlexNet model to identify dairy cattle, achieving 96.7% accuracy from 105 side-view images. However, most deep learning-based approaches focus on extracting spatial and semantic features from images. Hence, important temporal information, usually influenced by cattle motion or posture change, is mostly ignored.

As cattle videos are comprised of a large number of sequential images, they contain both spatial and temporal information such as back postures, kinematic gait parameters and walking behavior (Guo et al., 2021; Xue et al., 2021). To make the most of the video semantic contents, sequence learning models such as LSTM (Long Short Term Memory) and BiLSTM (Bidirectional long short-term memory) are required for cattle identification. LSTM is a commonly used model for solving sequential data problems, such as video recognition and text classification (Xie et al., 2019). LSTM can discover patterns and features when trained with back propagation through time (Karim et al., 2017). BiLSTM is a further development of LSTM that combines the forward hidden layer and the backward hidden layer (Li et al., 2020b). Unlike LSTM that only exploits historical information, BiLSTM can access both the preceding and succeeding sequence information. However, redundant information may hinder the acquisition of important features and decrease the accuracy of identification or recognition. LSTM and BiLSTM regard each frame equally and thus do not focus on pertinent information in features restricting identification accuracy.

The attention mechanism in deep learning has been widely used to further improve the performance of sequence learning tasks (Yang et al., 2019; Deng et al., 2020) and was proposed for discovering task-relevant features via attention weights (an informative sequence of weights). For example, Li et al. (2020a) and Song et al. (2018) used attention for weighting relevant frames for better understanding the action being carried out. In work of Piergiovanni et al. (2017), a series of temporal attention filters that weight frame-level features depending on their relevance for identifying actions, were learnt. Intuitively, a combination of the attention mechanism and BiLSTM can enhance the deep learning model's ability to capture important spatio-temporal features for cattle identification.

Here, we propose a deep learning based approach to cattle identification using video data. Especially, we propose an attention-based BiLSTM approach which uses more recognizable and relevant spatio-temporal features to improve the accuracy of cattle identification. As illustrated in **Figure 1**, the proposed approach consists of a CNN network for spatial feature extraction, a BiLSTM layer for mining correlations between different frames, and one attention layer for re-adjusting the spatio-temporal correlation for final cattle identification. More specifically, the CNN network extracts spatial features from each image in the videos, and then the BiLSTM layer accesses both the preceding and succeeding temporal features by combining a forward hidden layer and a backward hidden layer. After the BiLSTM layer, the attention layer priorities particular spatio-temporal features related to the cattle identification and helps to understand the video semantics. Finally, the weighted spatio-temporal features are fed to the softmax function for the identification of individual cattle.

The main contributions of our work can be summarized as follows: (1) We have introduced a deep learning framework for cattle identification using video data. By taking key beneficial features from both CNN and BiLSTM, our video based approach automatically learns spatio-temporal feature representations in the video data and is shown to outperform the image-frame based approach; (2) A simple and effective self-attention mechanism was employed to weight spatio-temporal features, maximizing the contribution of relevant spatio-temporal features and minimizing the influence of the irrelevant features for cattle identification. Experiment results showed that the attention mechanism can significantly enhance the effect of identification; (3) We have extensively compared the proposed approach with the state-of-the-art methods (Inception-V3, MLP, SimpleRNN, LSTM, and BiLSTM), and our results show that the proposed approach outperformed these methods; (4) The effects of two different attention mechanisms, namely, the additive attention mechanism and the multiplicative attention mechanism, were also investigated. Our experimental results favored the former one against the latter one; (5) We have also studied the influence of video length on accuracy, and showed increasing video length up to 30-frames enhanced identification performance.

The remainder of this paper is organized as follows: Section 2 reviews related works; Section 3 is the preliminary knowledge; Section 4 illustrates the proposed attention-based BiLSTM for

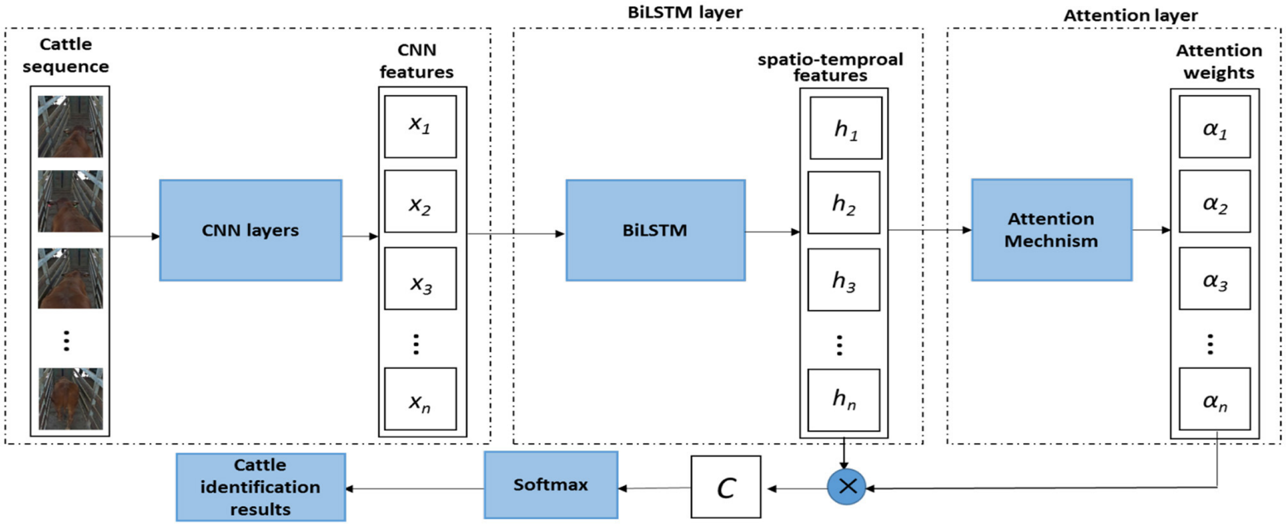

FIGURE 1 | The basic framework of the attention based BiLSTM for individual cattle identification.

cattle identification; Section 5 presents the experimental setup including dataset and evaluation methods; Experimental results and analysis are presented in Section 6; Further discussion on the proposed method and the experimental results are presented in Section 7; Conclusions are provided in Section 8.

2. RELATED WORKS

With the development of visual sensors and image analysis technologies, vision-based cattle identification as a non-contact approach is becoming feasible (Van Der Zande et al., 2021). For example, biometric and visual features, extracted from images of cattle muzzle, face, coat, torso, retinas and irises, have been shown to be helpful for identifying cattle (Jiang et al., 2019; Guzhva et al., 2021). Cai and Li (2013) and Kumar et al. (2017) presented a facial representation model of cattle based on extracted facial features, while Gaber et al. (2016) used the Weber Local Descriptor to extract robust features from cattle muzzle print images for cattle identification. Similarly, Kusakunniran et al. (2018) proposed an automatic cattle identification approach by fusing visual features extracted from muzzle images, while Andrew et al. (2016) utilized cattle coat patterns for identification. In addition, Zhao et al. (2019) extracted feature points of the cow body and matched them with the template dataset to identify cows. However, these manually selected and extracted features are impacted by cattle appearance change such as covering soil, animal dung or illumination, and camera viewpoints (Wurtz et al., 2019). Therefore, the handcrafted texture feature extraction and appearance-based feature representation techniques are not suitable for animal recognition in unconstrained environments.

Recently, deep learning approaches with powerful feature extraction and image representation abilities have been used in cattle identification (Qiao et al., 2019c; Shen et al., 2019a). For example, Zhao and He (2015) proposed a CNN network method for cow identification and Kumar et al. (2018) proposed a CNN based approach for identification of individual cattle by using primary muzzle point image pattern. Zin et al. (2018) trained a deep learning model based on back images to identify cows. However, the former approach ignored head and leg data which also contain useful identification information such as contour and texture features. Hu et al. (2020) extracted CNN features from cattle head, trunk and legs parts using side-view images, and then fused these features for individual cow based on Support Vector Machine (SVM) classifier. Their proposed method achieved 98.4% identification accuracy among 93 cows. In addition, Bezen et al. (2020) adopted Faster R-CNN to detect and identify cows, which achieved 93.7% accuracy. Despite the above progress, existing image frame-based approaches have not fully utilized temporal information such as walking posture or kinematic gait parameters.

Time-series data such as video, on the other hand, contain more temporal information and have been widely used to monitor cattle behaviors and welfare (Ordóñez and Roggen, 2016; Bahlo et al., 2019). For example, Van Hertem et al. (2014) developed an automated lameness scoring algorithm based on 3D-video recordings of cow gait. McPhee et al. (2017) developed a learning-based approach for assessing traits such as rump fat and muscle score. Nasirahmadi et al. (2017) implemented machine vision to detect cattle behavior. Inspired by this, temporal features extracted from videos are also gaining popularity for cattle identification. Andrew et al. (2017) demonstrated a video processing pipeline for cattle identification, which adopted a Long-term Recurrent Convolutional Network to classify cattle videos taken by the unmanned aerial vehicles. Karim et al. (2019a) proposed LSTM fully convolutional networks for time series classification. Okura et al. (2019) used

RGB-D videos of walking cows to extract two complementary features–gait (i.e., walking style) and texture (i.e., markings), which achieved 84.2% cow identification accuracy. In our recent work, Qiao et al. (2019b) proposed a beef cattle identification framework which uses image sequences, and combines CNN and BiLSTM for improving cattle identification accuracy (Qiao et al., 2020).

With the concept of "attention" gaining popularity in the deep learning field and its ability in guiding neural networks to learn more relevant features (Li et al., 2020b), attention-based LSTM models have been proposed for the task of video recognition to capture the most relevant temporal information in the video sequences (Wang et al., 2019). When the attention mechanism was first proposed in the field of computer vision, its purpose is to imitate the attention mechanism of human beings and give different weights to different parts of the image. Zeng et al. (2019) proposed attention-based LSTM with position context for aspect-level sentiment classification. Du et al. (2019) proposed convolution-based neural attention for sentiment classification. Guan et al. (2019) proposed a two-way LSTM model for attention enhancement for sentiment analysis. Xu et al. (2017) presented Multimodal-attention LSTM to encode and decode LSTM models with attention from different modalities and their related elements.

In this paper, we further develop and improve our previous work (Qiao et al., 2019b), providing a deep learning framework for beef cattle identification using video datasets. Similar to our previous work in Qiao et al. (2019b), the proposed framework synthesizes CNN and BiLSTM but also integrates the attention mechanism for mining key features to improve cattle identification performance. Here, a self-attention mechanism is introduced into the BiLSTM neural network model to improve the ability to capture key information, thereby further improving identification accuracy.

3. PRELIMINARY

3.1. LSTM and BiLSTM

LSTM is an extension of RNN (Recurrent Neural Networks) which is a popular model for solving sequential data problems (Karim et al., 2017). As depicted in **Figure 2**, LSTM uses a memory cell capable of representing the long-term dependencies in sequential data. The LSTM memory cell is composed of four gates (or units), namely, the input gate, the output gate, the forget gate, and the self-recurrent neuron. These gates are responsible for controlling the interactions among different memory units. Specifically, the input gate controls whether the input signal can modify the state of the memory cell or not. In contrast, the output gate controls whether it can modify the state of other memory cells or not. The forget gate can choose to forget (or remember) its previous status. At every time step t, given the input x, LSTM can choose to write, read or reset the memory cell through these three gates. This strategy helps LSTM to access and memorize information in many steps. The cell state c_t and the hidden value h_t of an LSTM are updated as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (3)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$

$$h_t = o_t \odot tanh(c_t) \quad (5)$$

where W_{h*}-like matrices are weight parameters that connect previous hidden states to each gate in an LSTM unit; similarly, W_{x*} indicates connections between current input and each gate, σ is the sigmoid function; tanh represents the hyperbolic tangent activation function; i, f, o, and c are the input gate, forget gate, output gate, and cell activation vectors, respectively; h is the hidden vector; b denotes bias vectors and matrix W is the

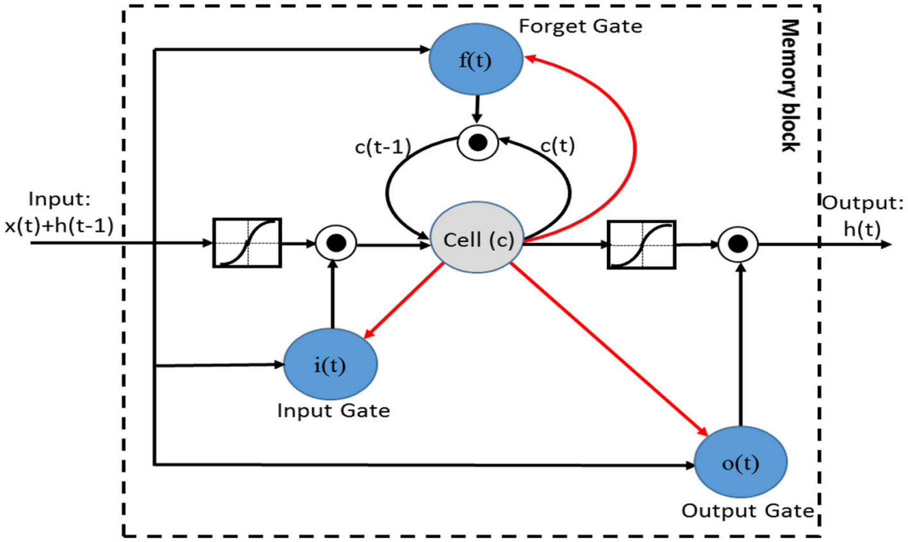

FIGURE 2 | LSTM cell.

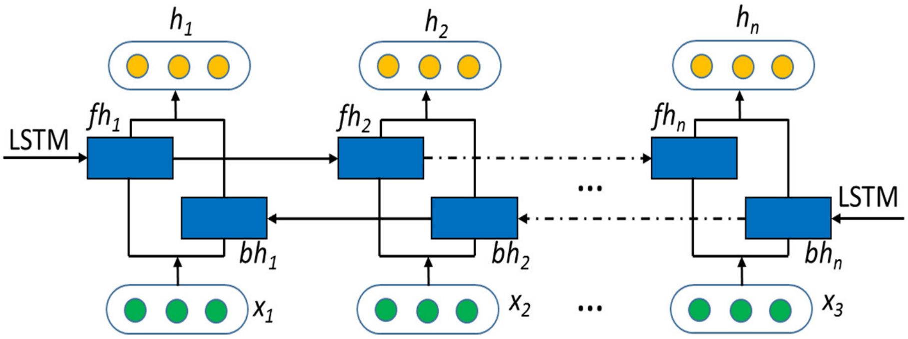

FIGURE 3 | BiLSTM network structure.

connection weight between two units; ⊙ represents the element-wise multiplication operation.

As a further development of LSTM, BiLSTM can access both the preceding and succeeding information through the forward hidden layer and the backward hidden layer. The BiLSTM model, as illustrated in **Figure 3**, consists of two independent LSTMs, which can sum up information from forward and backward directions of a sequence, and merge the information coming from the two directions.

3.2. Attention Mechanism

Self-attention, also known as intra-attention, is a special case of attention mechanism that only requires a single sequence to compute its representation, and has been successfully applied to many tasks, including machine translation and language understanding (Zhang et al., 2018). It provides the model with the ability to weight the features of single frames of the sequence differently, according to the similarity of the neighboring tokens (Zhang et al., 2018).

As shown in **Figure 4**, the self-attention mechanism has two components. (1) Attention weight calculation: the similarity score of each feature in the sequence is calculated, and then the similarity score is passed to a softmax function to generate the attention weights for each feature; (2) Weighted feature generation: the features obtained from BiLSTM are re-adjusted according to the corresponding attention weights.

4. MATERIALS AND METHODS

4.1. Data Acquisition

The cattle image dataset was collected at a Southern Queensland commercial feedlot in 2018 at three different times (induction, middle, and end point) on 20 March, 30 April, and 30 May, respectively. Our data were acquired when the cattle were walking along the race (path) from right to left in **Figure 5**. In our experiment, the left image of the rear view ZED camera was used; the image resolution was set to 1920×1080. A high frame acquisition rate (30 fps) was adopted to reduce the influences of motion blur during the herding process from the pen to the crush.

We extracted and saved the central part of the original 1920×1080 pixel image as the Region of Interest (ROI) to improve the system efficiency. This ROI was 401×506 pixels. In our experiment, a total of 363 cattle videos from 50 cattle were used, with each video containing 40-frame long spatio-temporal

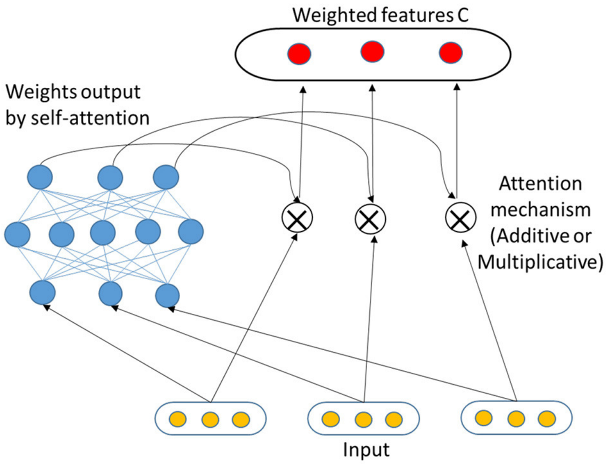

FIGURE 4 | General scheme of a self-attention mechanism applied to a sequence of features.

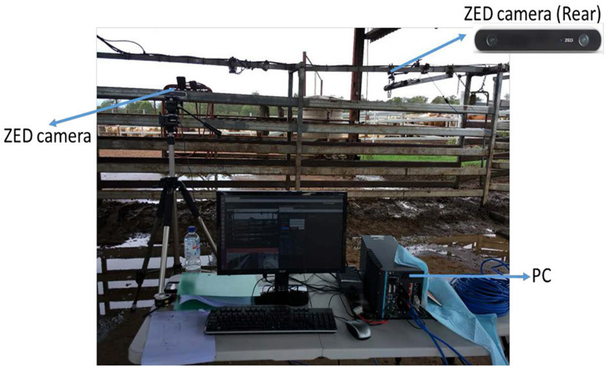

FIGURE 5 | Area for data acquisition, the race leading to the crush at a feedlot showing rear view camera clamped to overhead bar and the side view camera on tripod.

streams and each frame size as 401×506. For training, the true identities of cattle were manually recorded according to their ear tags. Our dataset is challenging for cattle identification considering different illumination conditions, animals' posture changes and complex background (including the cattle crush and ground).

4.2. The Proposed Attention-Based BiLSTM Approach

As illustrated in **Figure 6**, the proposed attention-based BiLSTM approach consists of four main important components:

- CNN feature extraction: For each image in the video, CNN features are extracted using pre-trained Inception-V3 network. These exacted features contain more spatial information, which represents cattle coat color or texture information. The extracted CNN features will be used in the next step.
- Spatio-temporal feature extraction: Based on the extracted CNN features, BiLSTM is used to model cattle video sequence considering the temporal evolution of the features for each time step. BiLSTM processes a sequence from forward and backward directions, and merges the information coming from the two directions. Capturing such information proves to be useful for cattle identification.
- Attention Mechanism: In order to focus on the important information and reduce the impact of non-relevant features, attention mechanism is applied to the extracted spatio-temporal features. By giving different weights to the spatio-temporal features, the attention mechanism can highlight the important information effectively. Thus spatio-temporal features related to the cattle identification are focused and more video semantics are explored. Those features with more relevant semantic relations to cattle distinguish ability are assigned with higher weights.
- Cattle identification using softmax layer: The weighted spatio-temporal features, a high-level representation of cattle video, are fed to the softmax layer for cattle identification.

For a given input cattle video containing N frame images $\{I_i\}_{i=1}^{N}$, firstly, features for each image in the video are extracted through a CNN network, which we denote as $\{X_i\}_{i=1}^{N} = \{x_1, x_2, \cdots, x_N\}$. These CNN features describe both the visual content and the spatial information of cattle images in the video. The CNN feature dimension of each image I can be labeled as d, which is determined by the feature extraction method. Each image in the video is converted into a d-dimension vector. After that, the extracted CNN features X are regarded as the input for the BiLSTM neutral network model aiming to obtain spatio-temporal features $H = \{h_1, h_2, \cdots, h_N\}$. Afterward, an attention layer is applied over each spatio-temporal feature $\{h_1, h_2, \cdots, h_N\}$. The attention mechanism can highlight important information from spatio-temporal features by setting different weights. Attention weights are obtained based on the similarity between hidden state representations of each spatio-temporal feature. Finally, the weighted spatio-temporal features are fed to the softmax layer to predict cattle ID.

4.2.1. CNN Feature Extraction

Inception-V3 is a popular model that can be used for image recognition and transfer learning (Szegedy et al., 2016), which is made up of building blocks including several layers such as convolutions, average pooling, max pooling, concats, dropouts, and fully connected layers. Therefore, in our work, we adopted Inception-V3 to extract CNN features from cattle images in each video. For a given image I_t at time t and learned parameter w, the

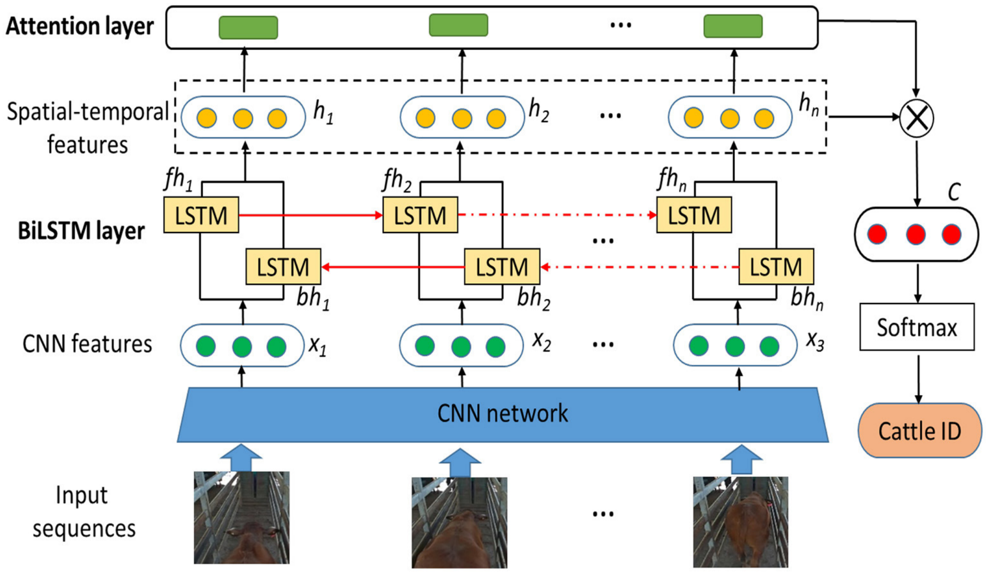

FIGURE 6 | The overall structure of proposed attention-based BiLSTM for cattle identification.

CNN features can be extracted by:

$$X(t) = f_N(\dots f_2(f_1(I_t;\, w_1);\, w_2)\dots),\, w_N) \tag{6}$$

where $f_1, \cdots, f_N$ are the corresponding layer functions. Each layer output is a CNN feature for image representation. In general, the low layers retain high spatial resolution whilst the high layers contain more semantic information. To maintain system performance and efficiency, the features extracted from the final pool layer in Inception-V3 were used in our work. Inception-V3 model with the pre-trained weights on the ImageNet dataset (Deng et al., 2009) was used to extract cattle features. Thus each image in the videos has a 2048-dimensional feature before being passed to the BiLSTM model.

4.2.2. Spatio-Temporal Feature Extraction Using BiLSTM

To extract key information from video content and to capture more video semantics, the visual aspects characterizing the object appearances as well as the motion present within the data should be considered. As such, after CNN features are extracted image by image from video, the next step is to extract spatio-temporal features.

More specifically, at each time t, given a video frame length n, the input CNN features are denoted as $[x_l, x_2, \cdots, x_N]$, the forward LSTM computes the hidden vector fh_t based on the previous hidden vector fh_{t-1} and the input CNN feature x_t. Meanwhile, the backward LSTM computes the hidden vector bh_t based on the opposite previous hidden vector bh_{t-1} and the input CNN feature x_t. Finally, the forward hidden vector fh_t and the backward hidden vector bh_t are concatenated to form the final hidden vector h_t:

$$fh_t = \sigma(W_{xi}x_t + W_{hi}fh_{t-1} + +b_i) \tag{7}$$

$$bh_t = \sigma(W_{xi}x_t + W_{hi}bh_{t-1} + +b_i) \tag{8}$$

$$h_t = [fh_t, bh_t] \tag{9}$$

BiLSTM, with the extracted CNN features as its input, is capable to learn and model each cattle's unique temporal characteristics. Considering that cattle can often walk back and forth, BiLSTM was applied to further extract spatio-temporal features of motion. In our work, the final hidden state h_t of BiLSTM was utilized as spatio-temporal feature to represent the cattle video data.

4.2.3. Attention Mechanism

For cattle identification, not all frames in a sequence are equally informative for the task. In particular, spatio-temporal features extracted from some image frames might be too quiet or noisy and thus contribute little to the identification clues (Zhang et al., 2020). In our work, a level of self-attention was implemented following the BiLSTM, which guides the model to focus on features that are relevant for the cattle identification task. For the aforementioned spatio-temporal features $[h_1,\, h_2,\, \cdots,\, h_N]$ of n-length sequence, the element $a_{t,t'}$ captures the similarity score between the hidden state representations h_t and $h_{t'}$ at t and t', respectively. Then the calculated similarity score $a_{t,t'}$ is passed to softmax function to obtain the attention weight $\alpha_{t,t'}$. The computation of attention weights are as follows:

$$\alpha_{t,t'} = \frac{exp(a_{t,t'})}{\sum_{t'=1}^{N} exp(a_{t,t'})} \tag{10}$$

The specific implementation of similarity score $a_{t,t'}$ mainly includes the additive attention mechanism and the multiplicative attention mechanism Shen et al. (2019b):

$$a_{t,t'} = \begin{cases} \sigma(W_a tanh(W_g h_t + W_{g'} h_{t'} + b_g)) & additive \\ \sigma(h_t W_a h_{t'} + b_a) & multiplicative \end{cases} \tag{11}$$

where σ is the element-wise sigmoid function; W_g and $W_{g'}$ are the weight matrices corresponding to the hidden states h_t and $h_{t'}$; W_a is the weight matrix corresponding to their non-linear combination; b_g and b_a are the bias vectors.

Finally, the weighted (attention-focused) spatio-temporal features c_t at timestamp t is obtained by attention weight $\alpha_{t,t'}$ and spatio-temporal features $h_t : c_t = \sum_{t'=1}^{N} \alpha_{t,t'} h_t$.

Essentially, with the help of self-attention, the weighted spatio-temporal features $c = [c_1, c_2, \cdots, c_N]$ are able to better represent the input video and enhance the cattle identification performance.

4.2.4. Attention-Based BiLSTM for Cattle Identification

In this work, cattle IDs are predicted using video data. Each cattle sequence has one unique ID. The weighted spatio-temporal features c generated by attention-based BiLSTM represent the cattle videos, and they are fed to a softmax classifier for the final cattle identification. The probability value p of the cattle IDs are obtained by the softmax classifier:

$$p = softmax(W_s c + b_s) \tag{12}$$

where p is the predicted result through the model; W_s is the weighted matrix; b_s is the bias. The output with the maximum value (class confidences, the value is between 0 and 1) is regarded as the cattle ID. If it matches the ground-truth (real ID), then it will be regarded as a true result. Otherwise, it is a false result.

In addition, considering its effectiveness and convenience in recognition tasks, categorical cross entropy loss function was used in our work, and its formula is as follows:

$$L_{loss} = \frac{-1}{M} \sum_{i=1}^{M} y_i log p_i + (1 - y_i) log(1 - p_i) \tag{13}$$

where M represents the number of training samples; y_i represents the ground-truth; p_i is the output prediction of the i-th sample.

The whole attention-based BiLSTM for cattle identification is illustrated in **Algorithm 1**.

5. EXPERIMENT SETUP

5.1. Network Training

Our methods were implemented using Keras (Chollet, 2015) on a DELL TOWER PC with GeForce GTX 1080 Ti GPU. Details of

```
Algorithm 1 : Attention-based BiLSTM for cattle identification
Input: Cattle video $V = \{I_i\}_{i=1}^{N}$; $N$ is the frame number.
Output: Cattle ID
    Step 1 CNN Feature Extraction: Through CNN network, converting image sequence $[I_1, I_2, \cdots, I_N]$ into corresponding CNN features $[x_1, x_2, \cdots, x_N]$.
    Step 2 Spatio-temporal feature extraction: Modeling CNN feature sentences using BiLSTM models, according to formulas (7)–(9), and learning the spatio-temporal features $[h_1, h_2, \cdots, h_N]$ for each image.
    Step 3 Attention mechanism–obtaining the attention weights: Through the formula (10)–(11), computing the attention weight $\alpha_{t,t'}$ for each spatio-temporal feature.
    Step 4 Attention mechanism–weighted feature calculation: The weighted feature $c$ is calculated through sum of spatio-temporal features and corresponding attention weights.
    Step 5 Cattle identification: Adopting the feature $c$ obtained in step 4 to represent cattle video, and it is fed to the softmax classifier for cattle identification (Equation 12). The output value $p$ with the maximum confidence is the predicted cattle ID by the proposed approach.
```

TABLE 1 | The experimental hardware.

Hardware	Type
CPU	Intel Xeon E5-2630 @ 2.20 GHz ×20
Memory	32GB
GPU	GeForce GTX 1080 Ti
Hard disk	1 TB

hardware information for the current experiment are provided in **Table 1**.

In the proposed attention-based BiLSTM approach, the final pool layer of Inception-V3 was used to extract CNN features. For each image, 2048 dimensional CNN features were obtained. Then all the CNN features for each frame in the videos were passed to BiLSTM model. The above was followed by the attention mechanism, which gives a different focus to the spatio-temporal features extracted by BiLSTM. The final obtained weighted spatio-temporal features were used to identify cattle. In addition, a regularizer $L2$ with the value of 10^{-4} was also applied on the kernel, bias and attention layers, respectively, to prevent over-fitting for the network training.

In our experiments, 288 videos of 50 cattle were used for training while the remaining 75 videos of the same 50 cattle were used for testing. For comparison, all methods were trained and tested on the same dataset. In our work, for network training, the initial learning rate was 10^{-5}, learning decay factor was 10^{-6}, batch size was 10 and loss function was "categorical cross-entropy."

5.2. Comparison With Other Methods

We evaluate and compare our proposed attention-based BiLSTM model with several state-of-the-art approaches — Inception-V3, Multilayer Perceptron (MLP), LSTM, SimpleRNN, LSTM and BiLSTM.

- Inception-V3: In the Inception-V3 method, images from training videos (i.e., 14,400 frames) are used to train network whilst the images from testing videos (i.e., 3,750 frames) are utilized for testing. The number of output nodes in the last layer is equal to the number of cattle (50 in our data).
- MLP: MLP is a class of feed-forward artificial neural network, which is usually composed of an input layer that receives the signal, an output layer that makes a decision or prediction about the input, and an arbitrary number of hidden layers (Li and Cao, 2019). MLP can model the correlation between those inputs and outputs, therefore it is often applied for supervised learning tasks. In our work, a two-layer MLP with 2,048 neurons per layer was used.
- SimpleRNN: SimpleRNN is a kind of RNN (Recurrent Neural Network), which calculates hidden vector sequences and output vector sequences through a linear transform and an activation function (Guo et al., 2019).
- LSTM: LSTM is a popular network for space-time data processing with strong abilities to learn and remember over long sequences of input data (Karim et al., 2019b). It makes use of the "gating" concept to update cell states. Each gate is a non-linear summation unit which controls the operation of the cell memory (Itakura et al., 2019). In our experiments, one-layer LSTM and two-layer LSTM are adopted, respectively. Each LSTM layer has 2,048 cells with a dropout rate of 0.5.
- BiLSTM: The BiLSTM consists of two independent LSTMs, which can sum up information from forward and backward directions of a sequence, and merge the information coming from the two directions (Li et al., 2020b). In our experiments, 1 BiLSTM layer with 2,048 cells was used.

5.3. Performance Evaluation

Precision-recall characteristics and F_1-score are widely used to evaluate recognition tasks (Qiao et al., 2019a). In order to evaluate performance, accuracy, precision, recall and F_1-score were used to evaluate the performance of cattle identification. Accuracy is the ratio of the number of correct predictions to the total number of input samples; precision shows the ability of the model to accurately identify targets; recall reflects the ability of the model to detect targets; F_1 score is a harmonic means of the precision and recall— a perfect system would return a result where both precision and recall have a value of one. All the above four measures range from 0 to 1, high value means the good predictive ability of the model.

Cases when the models correctly predicted the positive class (current ID) and the negative class (Non-current ID) are defined as TP (True Positive) and TN (True Negative), respectively. Incorrect model prediction of the positive (current ID) and negative classes (Non-current ID) are defined as FP (False Positive) and FN (False Negative). Based on the above, accuracy, precision, recall and F_1-score definitions are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (14)$$

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (15)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (16)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \quad (17)$$

6. CATTLE IDENTIFICATION RESULTS

6.1. Comparison of Different Methods

The experimental results of the proposed approach compared with other state-of-the-art methods are shown in **Table 2** (here additive attention mechanism is used). The proposed attention-based BiLSTM achieved an accuracy of 93.3%, a precision of 89.3%, a recall of 91.0%, and an F_1 score of 90.2%, values greater than those of Inception-V3, SimpleRNN, LSTM and BiLSTM. In particular, the precision and F_1 score of the attention-based BiLSTM is approximately 5% greater than those of BiLSTM (84.0% precision and 85.5% F_1 score) meaning the proposed attention mechanism is able to focus on key spatio-temporal features, thereby significantly improving cattle identification accuracy.

From **Table 2**, the performance of MLP (82.7% accuracy), SimpleRNN (86.7% accuracy), LSTM (89.3% accuracy), and BiLSTM (90.7% accuracy) using video is greater than the image-frame based approach Inception-V3 (80.0% accuracy) due to the fact that LSTM and BiLSTM learning useful temporal information such as the gait or walking behavior of cattle, which further enhances visual cattle identification performance. The experimental results illustrated that video based approach can extract and learn extra information (temporal) relevant to individual identification from video data.

In addition, the model with two LSTM layers achieved 86.7% accuracy which is lower than that of one LSTM layer (89.3%) as our cattle video dataset is limited so that one layer LSTM could transfer the time series data into highly "condensed" semantic information, while extra layers cause information loss during the transfer process and gradient vanishing in the training process.

TABLE 2 | Cattle identification performance comparison of different methods (%).

Methods	Accuracy	Precision	Recall	F_1	Time (s)
Inception-V3	80.0	67.1	75.0	70.8	-
MLP	82.7	75.2	79.0	77.0	10.2
SimpleRNN	86.7	80.0	83.0	81.5	4.9
LSTM (one layer)	89.3	80.8	86.0	83.3	13.1
LSTM (two layer)	86.7	77.3	83.0	80.0	13.6
BiLSTM	90.7	84.0	87.0	85.5	26.2
Attention-BiLSTM	**93.3**	**89.3**	**91.0**	**90.2**	28.7

The bold values are high performance values.

Furthermore, the runtime of attention-based BiLSTM approach and other methods for testing dataset were also reported in **Table 2**. Although the attention-based approach took more time than other approaches, it also had the best performance.

Our proposed approach can not only maintain the spatial information of visual features for cattle identification but also effectively pay attention to the most relevant spatio-temporal features.

6.2. Confusion Matrix of Cattle Identification

In order to further analyze the performance of attention-based BiLSTM for cattle identification, confusion matrices of the cattle recognition using the conventional BiLSTM and the proposed attention-based BiLSTM are presented in **Figure 7**. It can be seen that the performance of the proposed attention-based BiLSTM is better than that of BiLSTM (outliers in **Figure 7A** is fewer than that in **Figure 7B**). This is because the attention mechanism focuses on cattle identification related spatio-temporal features, which improves the distinguishing ability between different cattle.

The overall performance of the attention-based BiLSTM based beef cattle identification was favorable except for a few false identifications. **Figure 8** illustrates typical true and false cattle identification examples of the proposed attention-based BiLSTM approach. There are a few false cases in which cattle were partly covered by mud. In addition, for some videos, cattle were standing static or made little movement. For these cases, the cattle identification accuracy was not as high due to the lack of enough temporal information.

6.3. Effect of Different Attention Mechanisms

Additive and multiplicative are two main attention mechanisms are for cattle identification considered in this our work.

The performance comparisons over optimization epochs are shown in **Figure 9**. The blue and red solid lines in **Figures 9A,B** represent the reduction of loss and accuracies of the proposed approach with additive and multiplicative attention mechanism, respectively.

In **Figure 9A**, it can be seen that the losses of additive attention mechanism (red solid line) get close to 0 after 250 epochs' training, which is faster than that of multiplicative attention mechanism (blue solid line). Furthermore, the accuracy of additive attention mechanism (red solid line) in **Figure 9B** is greater than that of multiplicative attention mechanism (blue solid line). These further illustrate the advantages of the additive attention mechanism for operating speed and accuracy.

In addition, **Table 3** compares two different attention mechanisms (additive and multiplicative attention mechanism). It can be seen that the additive attention mechanism achieved

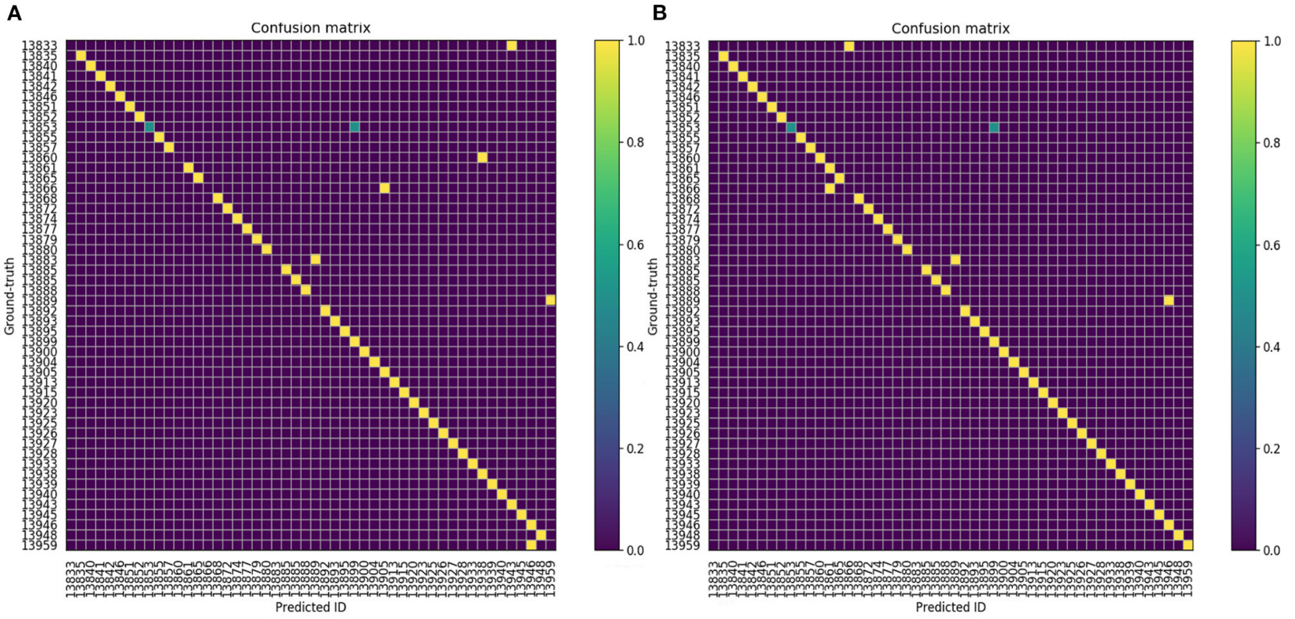

FIGURE 7 | Confusion matrix of cattle identifications. Each number indicates a cattle ID. The color indicates the similarity, yellow color means the highest similarity. **(A)** BiLSTM. **(B)** Attention-based BiLSTM.

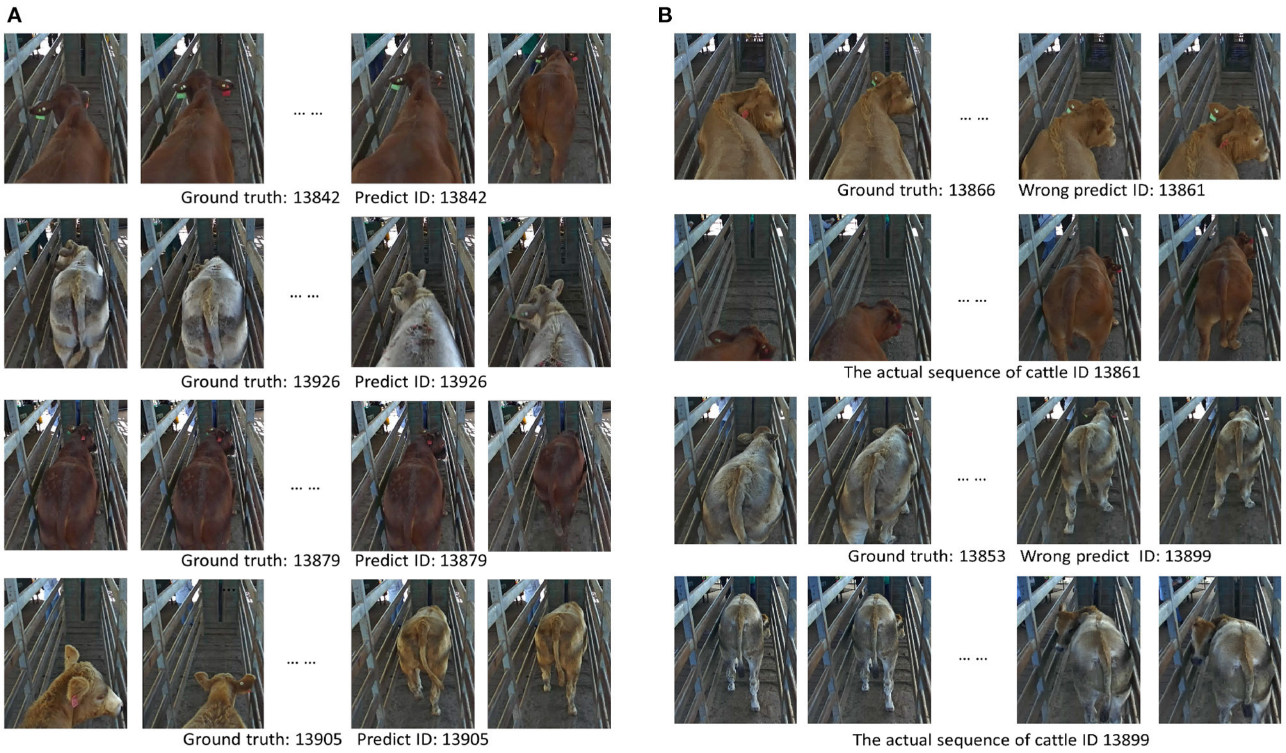

FIGURE 8 | Examples of attention based BiLSTM cattle identification results. The used video was 30-frame length. Note: for each video, only first two and the last two frames are displayed; some cattle were walking back and forth during data acquisition. **(A)** True identification examples. **(B)** False identification examples.

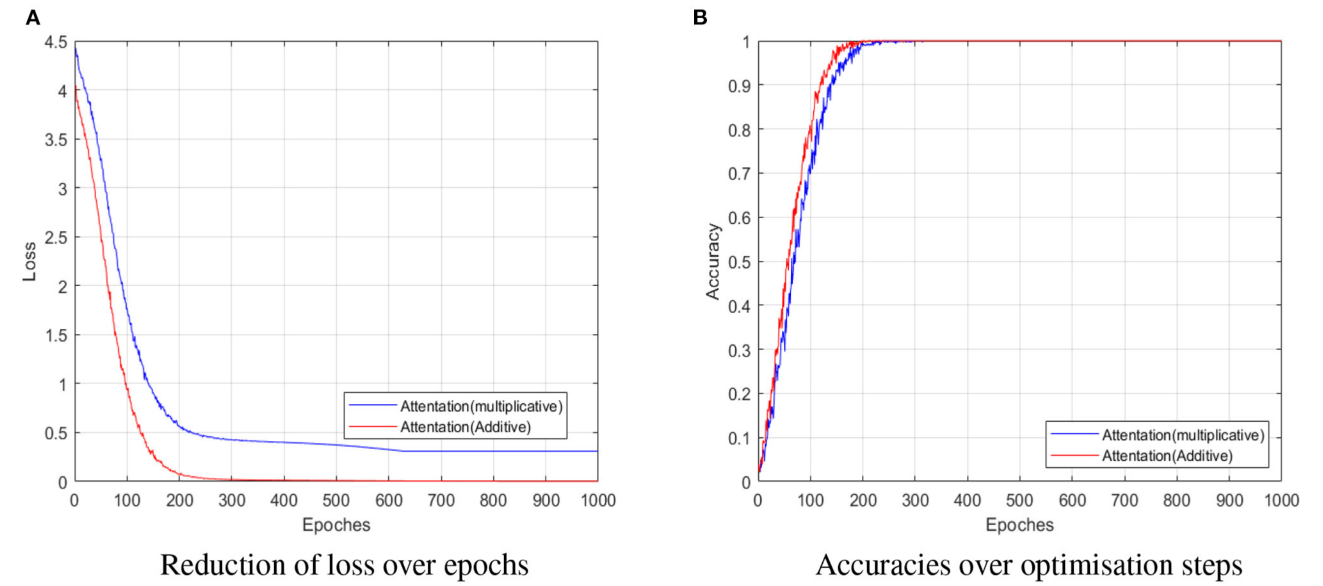

FIGURE 9 | Performance of two different attention mechanism on training dataset. Reduction of losses **(A)**, and accuracies **(B)** over optimization epochs. The blue and red solid lines in **(A,B)** represent reduction of loss and accuracies of the proposed approach with additive and multiplicative attention mechanism, respectively.

TABLE 3 | Comparison of different attention mechanisms (%).

Methods	Accuracy	Precision	Recall	F_1	Time (s)
Multiplicative	90.7	84.0	87.0	85.5	27.2
Additive	**93.3**	**89.3**	**91.0**	**90.2**	28.7

The bold values are high performance values.

93.3% accuracy, 89.3% precision, 90.2% F_1, and 91.0% recall, which outperformed those of the multiplicative attention mechanism (90.7% accuracy, 84.0% precision, 85.5% F_1, and 87.0% recall). A possible reason is that the additive attention mechanism assigns weights to spatio-temporal features more reasonably and retains more information than the latter one, although the multiplicative attention mechanism is faster and more memory-efficient due to optimized matrix multiplication.

7. DISCUSSIONS

7.1. Analysis of the Attention-Based BiLSTM for Cattle Identification

In our work, an attention-based BiLSTM approach for beef cattle identification using video data is proposed. The proposed approach leverages the strengths of CNN and BiLSTM, which is efficient in extracting spatio-temporal information (Qiao et al., 2019a), and modeling the hidden patterns or features in time-space data (Karim et al., 2019b). Meanwhile, the attention mechanism is employed to give different focus to the spatio-temporal features extracted by BiLSTM. The proposed approach captures more relevant and important spatio-temporal features, thereby improving the accuracy of cattle identification.

For our real beef cattle datasets, attention-based BiLSTM obtained better identification results than other state-of-the-art methods (Inception-V3, MLP, LSTM, and BiLSTM). As illustrated in **Table 2**, our proposed attention-based BiLSTM approach achieved 93.3% cattle identification accuracy, which is 2.6, 6.6, and 13.3% than that of BiLSTM, LSTM (two layer) and Inception-V3, respectively. The main reasons why the proposed attention-based BiLSTM works well are three-fold: (1) compared to LSTM, BiLSTM can accesses both the preceding and succeeding video information; hence, BiLSTM can more effectively learn the information of each frame in the video; (2) the attention mechanism can identify the influence of each spatio-temporal feature on the video and assign different attention weights to spatio-temproal features, thereby capturing the important components of the video semantics; (3) the combination of BiLSTM and attention mechanism makes the understanding of video semantics more accurate and improves the cattle identification ability.

However, we also observed false recognition cases (see **Figure 8B**) in the results of the proposed attention-based BiLSTM. Mud or dung on the body may be the cause of reduced identification accuracy. Another possible reason is that the CNN layers used here are not very deep due to the restriction of our data size. Increasing the layer number and training images could further improve identification accuracy. In addition, some cattle have limited motion which also leads to few temporal features captured for final identification.

7.2. The Influence of Video Length for Cattle Identification

The accuracies of the proposed method with respect to different video lengths (i.e., 10, 20, 30, and 40 frame length) are shown in

Figure 10. The greatest accuracy of attention-based BiLSTM was 93.3%, which outperformed BiLSTM (90.7%) and LSTM (89.3%).

In the cases of 10 to 30 frames, accuracy of both LSTM and BiLSTM improved with increased sequence length as more useful spatio-temporal features were extracted from longer videos. However, for the video-length between 30 and 40 frames, the performance of LSTM did not improve whilst the attention-BiLSTM decreased as no new information was captured by LSTM, and even some interference information (noise) was captured by the attention-BiLSTM.

7.3. Analysis of Running-Time

For real livestock farming applications, real-time image analysis and cattle identification are important. Therefore, we also discussed the usability of our algorithm in real-time implementation.

Our current attention-BiLSTM based cattle identification was tested on the Nvidia GTX 1080 Ti GPU-equipped computer. The proposed attention-BiLSTM approach could be further optimized to implement on the embedded computing boards such as Jetson XT2. Based on the off-line pre-trained model and Jetson XT2, real-time cattle identification no matter from the robotic platform or unmanned aerial vehicle (UAV) could be practically feasible.

7.4. The Annotation Techniques for Cattle Identification

In our current work, the whole image is used to extract visual features for cattle identification. However, the image background is not helpful to the cattle identification. If we only extract features from cattle body part, the identification performance would be further improved. The recent work of Psota et al. (2019) introduced an image space representation to represent body part locations and pairwise associations, which leverages the power of deep learning to detect the location and orientation of each animal. The proposed method achieved over 99% animal detection precision in group-housing environments. The annotation technique established by Psota et al. (2019) could be used to locate the cattle location and orientation, and the detected animal body parts could be used to extract the visual biometric features for the final cattle identification.

In our future work, we will further improve cattle identification accuracy using a larger and more-complex video dataset. Methods to merge animal detection and visual animal biometrics will also be pursued.

8. CONCLUSIONS

Identifying individual cattle is required for precision livestock management. Here, we use an attention-based BiLSTM to mine spatial and temporal cattle information from video for improving identification accuracy. The proposed approach consisted of CNN layers, a BiLSTM layer and an attention Layer. The visual features were extracted using Inception-V3 CNN network. After that, BiLSTM layer was employed to capture spatio-temporal features, so that more semantic information could be captured. Then the self-attention mechanism was incorporated to focus on the spatio-temporal features related to cattle identification. Finally, the weighted spatio-temporal features were fed to softmax for cattle identification. The proposed attention-based BiLSTM achieved 93.3% accuracy, which outperformed Inception-V3, MLP, SimpleRNN, LSTM and BiLSTM. In addition, the effects of two different types of attention mechanisms and influence of sequence length were also revealed. The additive attention mechanism outperformed the multiplicative attention mechanism, meanwhile identification accuracy was enhanced with increased sequence length when up to 30-frames. Overall, this research could provide some technical references for automatic cattle identification in the applications of precision livestock farming.

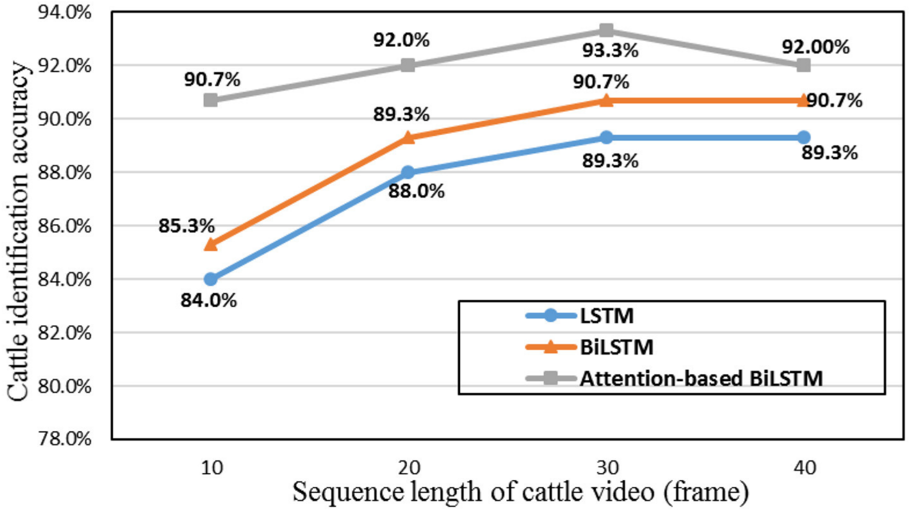

FIGURE 10 | Cattle identification accuracy w.r.t. sequence length. Cattle identification accuracy is changing with the used sequence length.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/qiao19881314/beefidentify.

ETHICS STATEMENT

The animal study was reviewed and approved by University of Sydney Animal Ethics Committees (AEC).

AUTHOR CONTRIBUTIONS

YQ: investigation, data curation, methodology (lead), formal analysis, and writing-original draft. CC: co-ordinating experiment in field, funding acquisition, and paper revising. SL: co-ordinating experiment in field, resources, and paper improvement. HK: methodology (supporting) and writing-review and editing. DS: initiated the idea, methodology (supporting), and writing-review and editing. SS: resources and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

The authors declare that this study received funding from Meat & Livestock Australia Donor Company (grant number P.PSH.0819). The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

The authors express their gratitude to Khalid Rafique, Javier Martinez, Amanda Doughty, Ashraful Islam, and Mike Reynolds for their help in experiment organization and data collection.

REFERENCES

Andrew, W., Greatwood, C., and Burghardt, T. (2017). "Visual localisation and individual identification of Holstein Friesian cattle via deep learning," in *Proc. IEEE International Conference on Computer Vision (ICCV)* (Venice), 22–29. doi: 10.1109/ICCVW.2017.336

Andrew, W., Hannuna, S., Campbell, N., and Burghardt, T. (2016). "Automatic individual holstein friesian cattle identification via selective local coat pattern matching in RGB-D imagery," in *2016 IEEE International Conference on Image Processing (ICIP)* (Phoenix, AZ: IEEE), 484–488. doi: 10.1109/ICIP.2016.7532404

Bahlo, C., Dahlhaus, P., Thompson, H., and Trotter, M. (2019). The role of interoperable data standards in precision livestock farming in extensive livestock systems: a review. *Comput. Electron. Agric*. 156, 459–466. doi: 10.1016/j.compag.2018.12.007

Berckmans, D. (2014). Precision livestock farming technologies for welfare management in intensive livestock systems. *Sci. Tech. Rev. Office Int. Epizooties* 33, 189–196. doi: 10.20506/rst.33.1.2273

Bezen, R., Edan, Y., and Halachmi, I. (2020). Computer vision system for measuring individual cow feed intake using rgb-d camera and deep learning algorithms. *Comput. Electron. Agric*. 172:105345. doi: 10.1016/j.compag.2020.105345

Cai, C., and Li, J. (2013). "Cattle face recognition using local binary pattern descriptor," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (Taiwan: IEEE), 1–4.doi: 10.1109/APSIPA.2013.6694369

Chollet, F. (2015). *Keras: The Python Deep Learning Library*. Available online at: https://keras.io

Dawkins, M. S. (2021). Does smart farming improve or damage animal welfare?: technology and what animals want. *Front. Anim. Sci*. 2:736536. doi: 10.3389/fanim.2021.736536

de Lima Weber, F., de Moraes Weber, V. A., Menezes, G. V., Junior, A., d,. S. O., Alves, D. A., et al. (2020). Recognition of pantaneira cattle breed using computer vision and convolutional neural networks. *Comput. Electron. Agric*. 175:105548. doi: 10.1016/j.compag.2020.105548

Deng, J., Cheng, L., and Wang, Z. (2020). Self-attention-based bigru and capsule network for named entity recognition. *arXiv preprint arXiv:2002.00735*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL: IEEE), 248–255. doi: 10.1109/CVPR.2009.5206848

Du, J., Gui, L., He, Y., Xu, R., and Wang, X. (2019). Convolution-based neural attention with applications to sentiment classification. *IEEE Access* 7, 27983–27992. doi: 10.1109/ACCESS.2019.2900335

Gaber, T., Tharwat, A., Hassanien, A. E., and Snasel, V. (2016). Biometric cattle identification approach based on weber's local descriptor and adaboost classifier. *Comput. Electron. Agric*. 122, 55–66. doi: 10.1016/j.compag.2015.12.022

Guan, P., Li, B., Lv, X., and Zhou, J. (2019). Attention enhanced bi-directional LSTM for sentiment analysis. *J. Chin. Inform. Proc*. 33, 105–111. doi: 10.22323/1.300.0014

Guo, X., Zhang, H., Yang, H., Xu, L., and Ye, Z. (2019). A single attention-based combination of CNN and RNN for relation classification. *IEEE Access* 7:1246712475. doi: 10.1109/ACCESS.2019.2891770

Guo, Y., Qiao, Y., Sukkarieh, S., Chai, L., and He, D. (2021). Bigru-attention based cow behavior classification using video data for precision livestock farming. *Trans. ASABE*. doi: 10.13031/trans.14658

Guzhva, O., Siegford, J., and Kolstrup, C. L. (2021). The Hitchhiker's guide to integration of social and ethical awareness in precision livestock farming research. *Front. Anim. Sci*. 2:725710. doi: 10.3389/fanim.2021.725710

Hu, H., Dai, B., Shen, W., Wei, X., Sun, J., Li, R., et al. (2020). Cow identification based on fusion of deep parts features. *Biosyst. Eng*. 192, 245–256. doi: 10.1016/j.biosystemseng.2020.02.001

Itakura, K., Saito, Y., Suzuki, T., Kondo, N., and Hosoi, F. (2019). Classification of soymilk and tofu with diffuse reflection light using a deep learning technique. *Agriengineering* 1, 235–245. doi: 10.3390/agriengineering1020017

Jiang, B., Wu, Q., Yin, X., Wu, D., Song, H., and He, D. (2019). Flyolov3 deep learning for key parts of dairy cow body detection. *Comput. Electron. Agric*. 166:104982. doi: 10.1016/j.compag.2019.104982

Karim, F., Majumdar, S., and Darabi, H. (2019a). Insights into LSTM fully convolutional networks for time series classification. *IEEE Access* 7, 67718–67725. doi: 10.1109/ACCESS.2019.2916828

Karim, F., Majumdar, S., Darabi, H., and Chen, S. (2017). Lstm fully convolutional networks for time series classification. *IEEE Access* 6, 1662–1669. doi: 10.1109/ACCESS.2017.2779939

Karim, F., Majumdar, S., Darabi, H., and Harford, S. (2019b). Multivariate LSTM-fcns for time series classification. *Neural Netw*. 116, 237–245. doi: 10.1016/j.neunet.2019.04.014

Kumar, S., Pandey, A., Satwik, K. S. R., Kumar, S., Singh, S. K., Singh, A. K., et al. (2018). Deep learning framework for recognition of cattle using muzzle point image pattern. *Measurement* 116, 1–17. doi: 10.1016/j.measurement.2017.10.064

Kumar, S., Singh, S. K., Singh, R., and Singh, A. K. (2017). "Recognition of cattle using face images," in *Animal Biometrics* (Singapore: Springer), 79–110. doi: 10.1007/978-981-10-7956-6_3

Kusakunniran, W., Wiratsudakul, A., Chuachan, U., Kanchanapreechakorn, S., and Imaromkul, T. (2018). "Automatic cattle identification based on fusion of texture features extracted from muzzle images," in *2018 IEEE*

International Conference on Industrial Technology (ICIT) (Lyon: IEEE), 1484–1489. doi: 10.1109/ICIT.2018.8352400

Li, J., Liu, X., Zhang, W., Zhang, M., Song, J., and Sebe, N. (2020a). Spatio-temporal attention networks for action recognition and detection. *IEEE Trans. Multimedia*. 22, 2990–3001. doi: 10.1109/TMM.2020.2965434

Li, W., Qi, F., Tang, M., and Yu, Z. (2020b). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*. 387, 63–77. doi: 10.1016/j.neucom.2020.01.006

Li, Y., and Cao, W. (2019). An extended multilayer perceptron model using reduced geometric algebra. *IEEE Access* 7, 129815–129823. doi: 10.1109/ACCESS.2019.2940217

McPhee, M., Walmsley, B., Skinner, B., Littler, B., Siddell, J., Cafe, L., et al. (2017). Live animal assessments of rump fat and muscle score in angus cows and steers using 3-dimensional imaging. *J. Anim. Sci*. 95, 1847–1857. doi: 10.2527/jas.2016.1292

Nasirahmadi, A., Edwards, S. A., and Sturm, B. (2017). Implementation of machine vision for detecting behaviour of cattle and pigs. *Livestock Sci*. 202, 25–38. doi: 10.1016/j.livsci.2017.05.014

Okura, F., Ikuma, S., Makihara, Y., Muramatsu, D., Nakada, K., and Yagi, Y. (2019). RGB-D video-based individual identification of dairy cows using gait and texture analyses. *Comput. Electron. Agric*. 165:104944. doi: 10.1016/j.compag.2019.104944

Ordóñez, F., and Roggen, D. (2016). Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16:115. doi: 10.3390/s16010115

Piergiovanni, A., Fan, C., and Ryoo, M. S. (2017). "Learning latent subevents in activity videos using temporal attention filters," in *Thirty-First AAAI Conference on Artificial Intelligence*, Vol. 31 (San Francisco, CA), 4247–4254.

Psota, E. T., Mittek, M., Pérez, L. C., Schmidt, T., and Mote, B. (2019). Multi-pig part detection and association with a fully-convolutional network. *Sensors* 19:852. doi: 10.3390/s19040852

Qiao, Y., Cappelle, C., Ruichek, Y., and Yang, T. (2019a). Convnet and LSH-based visual localization using localized sequence matching. *Sensors* 19:2439. doi: 10.3390/s19112439

Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., et al. (2021). Intelligent perception for cattle monitoring: a review for cattle identification, body condition score evaluation, and weight estimation. *Comput. Electron. Agric*. 185:106143. doi: 10.1016/j.compag.2021.106143

Qiao, Y., Su, D., Kong, H., Sukkarieh, S., Lomax, S., and Clark, C. (2019b). Individual cattle identification using a deep learning based framework. *IFAC Pap. Online* 52, 318–323. doi: 10.1016/j.ifacol.2019.12.558

Qiao, Y., Su, D., Kong, H., Sukkarieh, S., Lomax, S., and 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)* (Hong Kong: IEEE), 967–972. doi: 10.1109/CASE48305.2020.9217026

Qiao, Y., Truman, M., and Sukkarieh, S. (2019c). Cattle segmentation and contour extraction based on mask R-CNN for precision livestock farming. *Comput. Electron. Agric*. 165:104958. doi: 10.1016/j.compag.2019.104958

Rotter, P. (2008). A framework for assessing RFID system security and privacy risks. *IEEE Pervas. Comput*. 2, 70–77. doi: 10.1109/MPRV.2008.22

Shen, W., Hu, H., Dai, B., Wei, X., Sun, J., Jiang, L., et al. (2019a). Individual identification of dairy cows based on convolutional neural networks. *Multimedia Tools Appl*. 79, 14711–14724. doi: 10.1007/s11042-019-7344-7

Shen, Y., Fang, Z., Gao, Y., Xiong, N., Zhong, C., and Tang, X. (2019b). Coronary arteries segmentation based on 3D FCN with attention gate and level set function. *IEEE Access* 7, 42826–42835. doi: 10.1109/ACCESS.2019.2908039

Song, S., Lan, C., Xing, J., Zeng, W., and Liu, J. (2018). Spatio-temporal attention-based LSTM networks for 3D action recognition and detection. *IEEE Trans. Image Process*. 27, 3459–3471. doi: 10.1109/TIP.2018.2818328

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 2818–2826. doi: 10.1109/CVPR.2016.308

Van Der Zande, L., Guzhva, O., Rodenburg, T. B., et al. (2021). Individual detection and tracking of group housed pigs in their home pen using computer vision. *Front. Anim. Sci.* 2:10. doi: 10.3389/fanim.2021.669312

Van Hertem, T., Viazzi, S., Steensels, M., Maltz, E., Antler, A., Alchanatis, V., et al. (2014). Automatic lameness detection based on consecutive 3d-video recordings. *Biosyst. Eng*. 119, 108–116. doi: 10.1016/j.biosystemseng.2014.01.009

Wang, Q., Luo, H., Ye, L., Men, A., Zhao, F., Huang, Y., et al. (2019). Pedestrian heading estimation based on spatial transformer networks and hierarchical LSTM. *IEEE Access* 7, 162309–162322. doi: 10.1109/ACCESS.2019.2950728

Wurtz, K., Camerlink, I., D'Eath, R. B., Fernández, A. P., Norton, T., Steibel, J., et al. (2019). Recording behaviour of indoor-housed farm animals automatically using machine vision technology: a systematic review. *PLoS ONE* 14:e0226669. doi: 10.1371/journal.pone.0226669

Xie, J., Chen, B., Gu, X., Liang, F., and Xu, X. (2019). Self-attention-based BiLSTM model for short text fine-grained sentiment classification. *IEEE Access* 7, 180558–180570. doi: 10.1109/ACCESS.2019.2957510

Xu, J., Yao, T., Zhang, Y., and Mei, T. (2017). "Learning multimodal attention LSTM networks for video captioning," in *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, CA), 537–545. doi: 10.1145/3123266.3123448

Xue, T., Qiao, Y., Kong, H., Su, D., Pan, S., Rafique, K., et al. (2021). One-shot learning-based animal video segmentation. *IEEE Trans. Indus. Inform*. doi: 10.1109/TII.2021.3117020

Yang, B., Li, J., Wong, D. F., Chao, L. S., Wang, X., and Tu, Z. (2019). "Context-aware self-attention networks," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), 387–394. doi: 10.1609/aaai.v33i01.3301387

Zeng, J., Ma, X., and Zhou, K. (2019). Enhancing attention-based LSTM with position context for aspect-level sentiment classification. *IEEE Access* 7, 20462–20471. doi: 10.1109/ACCESS.2019.2893806

Zhang, C., Cui, C., Gao, S., Nie, X., Xu, W., Yang, L., et al. (2018). Multi-gram CNN-based self-attention model for relation classification. *IEEE Access* 7, 5343–5357. doi: 10.1109/ACCESS.2018.2888508

Zhang, J., Xie, Z., Sun, J., Zou, X., and Wang, J. (2020). A cascaded r-CNN with multiscale attention and imbalanced samples for traffic sign detection. *IEEE Access* 8, 29742–29754. doi: 10.1109/ACCESS.2020.2972338

Zhao, K., and He, D. (2015). Recognition of individual dairy cattle based on convolutional neural networks. *Trans. Chin. Soc. Agric. Eng*. 31, 181–187. doi: 10.3969/j.issn.1002-6819.2015.05.026

Zhao, K., Jin, X., Ji, J., Wang, J., Ma, H., and Zhu, X. (2019). Individual identification of Holstein dairy cows based on detecting and matching feature points in body images. *Biosyst. Eng*. 181, 128–139. doi: 10.1016/j.biosystemseng.2019.03.004

Zin, T. T., Phyo, C. N., Tin, P., Hama, H., and Kobayashi, I. (2018). "Image technology based cow identification system using deep learning," in *Proceedings of the International MultiConference of Engineers and Computer Scientists* (Hong Kong), 236–247.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.