Visual Localisation and Individual Identification of Holstein Friesian Cattle via Deep Learning

William Andrew   Colin Greatwood   Tilo Burghardt

Department of Computer Science
University of Bristol

{will.andrew,colin.greatwood}@bristol.ac.uk, tilo@cs.bris.ac.uk

Abstract

In this paper, we demonstrate that computer vision pipelines utilising deep neural architectures are well-suited to perform automated Holstein Friesian cattle detection as well as individual identification in agriculturally relevant setups. To the best of our knowledge, this work is the first to apply deep learning to the task of automated visual bovine identification. We show that off-the-shelf networks can perform end-to-end identification of individuals in top-down still imagery acquired from fixed cameras. We then introduce a video processing pipeline composed of standard components to efficiently process dynamic herd footage filmed by Unmanned Aerial Vehicles (UAVs). We report on these setups, as well as the context, training and evaluation of their components. We publish alongside new datasets: FriesianCattle2017 of in-barn top-down imagery, and AerialCattle2017 of outdoor cattle footage filmed by a DJI Inspire MkI UAV. We show that Friesian cattle detection and localisation can be performed robustly with an accuracy of 99.3% on this data. We evaluate individual identification exploiting coat uniqueness on 940 RGB stills taken after milking in-barn (89 individuals, accuracy = 86.1%). We also evaluate identification via a video processing pipeline on 46,430 frames originating from 34 clips (approx. 20 s length each) of UAV footage taken during grazing (23 individuals, accuracy = 98.1%). These tests suggest that, particularly when videoing small herds in uncluttered environments, an application of marker-less Friesian cattle identification is not only feasible using standard deep learning components – it appears robust enough to assist existing tagging methods.

1. Introduction and Motivation

This paper aims at providing a proof of concept that robust individual Holstein Friesian cattle identification can occur *automatically* and *non-intrusively* using computer vision pipelines fuelled by standard architectures utilising deep neural networks. We will explore practically relevant
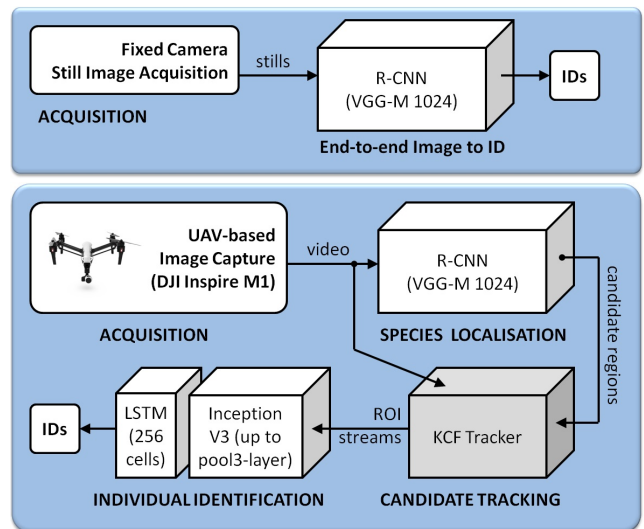

Figure 1: **Proposed Component Pipelines.** *(top)* Off-the-shelf baseline layout utilising still image input and comprising a VGG-M 1024 end-to-end R-CNN trained for individual Friesian cattle identification. *(bottom)* Extended aerial video monitoring pipeline with separation of species detection and individual identification components, added KCF tracking unit for fast trajectory extraction, and enhanced temporally-integrating individual identification component combining an Inception V3 network feeding into a standard LSTM unit to utilise complementary information as revealed across several frames in tracked RoI streams.

imaging setups using fixed and mobile camera platforms, that is, top-down observations in-barn, as well as airborne monitoring of freely grazing herds, respectively (see Fig. 1).

Holstein Friesian cattle are the highest milk-yielding bovine type [57]; they are economically important and especially prevalent within the UK [44]. Identification and traceability of these cattle is not only required by export and consumer demands [6], but in fact many countries have introduced legally mandatory frameworks [47] re-

IEEE
computer
society

volving around national databases in conjunction with ear-tagging [30, 7, 54]. Despite the heavy reliance of bovine identification upon this tagging approach, its effectiveness – as opposed to branding, tattooing [6] or electronic solutions [51] – has been called into question on numerous occasions, largely since ear-tags are subject to loss and physical damage [41]. In addition, animal welfare concerns have been raised regarding ear-tagging [16, 60]. Therefore, the automated and non-intrusive nature of visual bovine identification based on coat patterns promises both to aid efficiency within the farm and to contribute to animal welfare.

2. Related Work

2.1. Biometric Bovine Identification

The problem of automated biometric bovine identification has been well-studied over time. Approaches may be separated into three categories: those that utilise cattle muzzle patterns, rarer systems that employ retinal, facial or body scans, and those that exploit coat pattern characteristics.

Cattle muzzle patterns were first introduced as an individually-unique, dermatoglyphic trait in 1922 [48]. This seminal work was extended in 1993 where significant differences in cattle dermatoglyphics across breeds were discovered [8]. Muzzle patterns form the foundation of much research into semi-automated bovine identification methods [35, 37]. In many cases, feature-description algorithms such as Scale-Invariant Feature Transform (SIFT) [39] or Speeded-Up Robust Features (SURF) [4] are employed. In particular, Noviyanto et al. utilise SIFT on 160 manually-acquired muzzle prints of 20 individuals after applying ink to the cattle's muzzle [46], similarly to Minagawa et al. [42]. These semi-automated solutions are improved by approaches operating on muzzle images alone [36, 58, 45, 25]. As an example, Awad et al. combine SIFT and Random Sample Consensus (RANSAC) for feature extraction and filtering to achieve 93.3% identification accuracy on muzzle images [19].

Alternative approaches to bovine identification include constructing cattle facial representations based on Local Binary Pattern (LBP) texture features [9], utilising 3D visual appearances [3] and retinal imaging technology [1]. Note that advanced non-biometric computerised vision schemes for identification exist too, including work that takes advantage of the European cattle ear-tagging mandate [47] by applying text and character recognition in order to match tagged individuals against respective Bovine Identification Documents [59].

2.1.1 Coat Pattern Identification of Friesian Cattle

Holstein Friesian cattle are distinctive for exhibiting individually-unique black & white (or brown & white) coat patterns, patches and markings over their bodies (see Fig. 2

Figure 2: **UAV Acquisition of a Friesian Cattle Herd.** Representative frame from AerialCattle2017 captured by a DJI Inspire MkI UAV filming at 5 meters above ground. Individual cows are resolved at approximately 100×300 pixels within frames resolved at 3840×2160 pixels captured at a frame-rate of 24 fps. The animals' distinctive black and white coat patterns are clearly resolved and are used as an individually-characteristic biometric entity in this work.

for examples). Building on early systems that used Principal Component Analysis (PCA) and SIFT on image sequences [40], previous work has shown that dorsal patterns *alone* form sufficiently complex visual alignments for robust individual identification across a small herd [2]. This was achieved by extracting Affine SIFT (ASIFT) feature descriptors [43] from RGB-D-segmented coat patterns and learning those properties that contribute beneficially towards identification via a Radial Basis Function Support Vector Machine (RBF-SVM). However, this approach comes with a high computational burden due to extensive computations required for ASIFT feature calculation and matching, which places limits on the practical use of the framework.

2.2. Key Computational Methodologies

2.2.1 Deep Neural Architectures

Recent improvements in deep Convolutional Neural Networks (CNNs) have seen them outperform traditional computer vision techniques in object detection [53, 22] as well as image classification tasks across benchmark datasets [34]. Regional Convolutional Neural Networks (R-CNNs) combine these two tasks; candidate object locations are determined and subsequently classified [23]. Yet, in its original form, R-CNNs are computationally expensive to train and evaluate. With the introduction of sharing convolutions across proposals in Fast R-CNN and SPP-net [21, 26], significant improvements to efficiency were made, although region proposal computation remained the bottleneck. Circumventing this problem, Ren et al. propose the addition of a Region Proposal Network (RPN), which shares convolutional features with the detection net-

work leading to Faster-RCNN [49]. Furthermore, Region-based Fully Convolutional Networks (R-FCNs) [12] – based on fully convolutional architectures, such as FCN [38] – go further and avoid the per-proposal evaluation of Fast and Faster R-CNN. This change results in a significant speedup over Faster R-CNN on the PASCAL VOC datasets [12].

2.2.2 Recurrent and Memory-bearing Networks

Recurrent Neural Networks (RNNs) introduce the notion of memory across multiple evaluations of a network by additionally outputting information to the subsequent iteration. This renders them useful for tasks such as image sequence captioning where the processing of temporal or linked information requires retention.

Despite their intuitive simplicity, basic RNNs are unable to learn long-term dependencies [5] via traditional training methods such as Back-Propagation Through Time (BPTT) [61] or Real-Time Recurrent Learning (RTRL) [50]. This is due to the temporal progression of back-propagated error values being exponentially dependent on the size of weights [28]. First introduced in 1997, Long Short-Term Memory networks (LSTMs) [29] eradicate this problem by design. The architecture of LSTM cells or units enforce constant error flow, therefore preventing back-propagated errors from exploding or vanishing [28].

Following the success of LSTM networks, extensions or variations upon the original cell architecture have included allowing gate layers to examine the cell state [20], arranging networks of LSTM cells in a multidimensional grid [33], introducing a depth gate to connect memory cells of adjacent layers (Depth-Gated LSTM) [62], and many more. Proposed in 2014, Gated Recurrent Units (GRUs) [11] have been gaining popularity. They combine LSTM forget and input gates together into a single "update" gate as well as other modifications resulting in a simpler model overall. However, despite all these efforts, Greff et al. [24] find that no variants significantly outperform the original LSTM architecture over large-scale analysis of eight LSTM variants covering three popular machine learning tasks. Furthermore, Jozefowicz et al. [32] identify a RNN architecture that outperforms both LSTM and GRU units at only a small proportion of evaluated tasks. Thus, the original LSTM architecture is used in this work.

2.3. Paper Outline

The remainder of this paper is organised as follows: Sections 3 and 4 describe the employed approaches to cattle detection and identification, respectively. Section 5 provides details about the acquisition and pre-processing of the two datasets: FriesianCattle2017 and AerialCattle2017. Results and evaluations are then given in Section 6, whilst concluding remarks and details of future work are provided in Section 7.

3. Species Detection and Localisation

A first principal goal is to detect and locate Friesian cattle in images obtained from a top-down or aerial standpoint (see Fig. 3). The deep network utilised to address this problem is the R-CNN adaptation of the VGG CNN M 1024 network published as part of several other network architecture proposals [10]. The core architecture consists of 5 stacked convolutional layers, which are shared with the region proposal network – plus two fully connected layers. Instead of training from scratch, weights were initialised with a model trained on the ImageNet database [13] supplied with the Faster-RCNN Python implementation [49].

The object detection and localisation task produces image Regions of Interest (RoIs) or proposals in the form of bounding boxes. Proposals can then be passed to a subsequent process such as a tracking algorithm or individual classifier. Operating on single images, forward propagation on an image yields a set of n bounding box rectangles $\{bbox_{pred}^1, bbox_{pred}^2, ..., bbox_{pred}^n\}$ defined spatially by $bbox_{pred}^i = ((x_1^i, y_1^i), (x_2^i, y_2^i))$ for all $0 < i \leq n$. Each predicted object RoI is also associated with scores that reflect membership of each of the target class(es) such as cow and $background$ (see Fig. 3). By placing a confidence threshold on the scores for the trained target class cow, predicted RoIs can be retained for further processing or discarded appropriately.
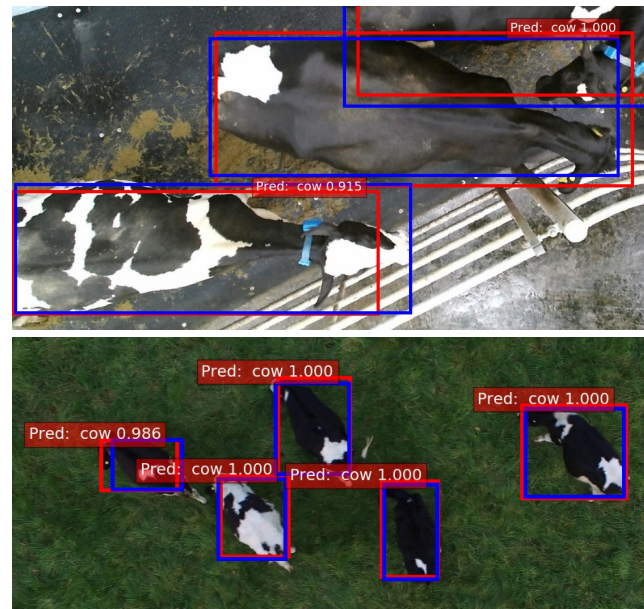

Figure 3: **Species Detection and Localisation.** Examples of species detection and localisation for the *(top)* Friesian-Cattle2017 indoor and *(bottom)* AerialCattle2017 outdoor datasets using a R-CNN. *(blue)* Ground truth bounding boxes for object class cow and *(red)* predicted cow object bounding boxes with class membership score values.

4. Individual Identification

4.1. Single Frame Individual Identification

In order to produce the simplest possible individual identification pipeline using R-CNNs operating on single images, the $classes$ set is extended to include all m individuals of a particular cattle population explicitly. That is, $classes = \{background, cow_1, cow_2, ..., cow_m\}$ where $|classes| = m + 1$ to include the negative class case also ($background$). In this way, the tasks of species detection and individual identification for trained individuals are incorporated naturally within a single end-to-end process (see Fig. 1, top). Whilst this setup is the most simple to build and replicate, it does not allow for an effective and efficient handling and processing of video.

4.2. Video-based LRCN Identification

Video – as opposed to a single still image of a scene, event or environment – intrinsically provides an additional (temporal) dimension of information relevant to individual identification. We propose to incorporate information from subsequent frames into the identification estimate to benefit from complementary information revealed progressively.

In the vast majority of cases, an individual cow can be tracked well within herd videos using the standard Kernelised Correlation Filter (KCF) tracking algorithm [27] given a good initialisation RoI yielded from the aforementioned species localisation component. This is since on average, cattle walk relatively slowly – D'Hour et al. find the average Holstein walking speed to be 1.37 m/s over distances of 5.6 and 3.2 km for two trials [14]. Additionally, cattle are unlikely to suddenly start walking backwards or change their heading by more than $\pm30°$. Thus, if cow_x is present in frame f_i, it is also likely to be present in f_{i+1} (given a sufficiently frequent frame-rate in source footage). These factors combined with the fact that UAV-captured source footage will exhibit positional and rotational vari-
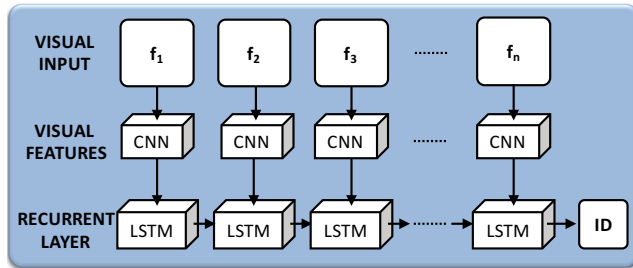

Figure 4: **Recurrent Convolutional Architecture.** Unrolled identification refinement pipeline for an input video based on the LRCN architecture [15]. Visual features for input video frames $\{f_1, f_2, ..., f_n\}$ are extracted via a CNN for input into a LSTM layer ultimately yielding a ID prediction.

ation due to winds, GPS inaccuracy, etc. contribute towards variation in viewpoint, object configuration and/or scale. However importantly, this often reveals the presence of further salient visual features useful for identification purposes. Such continual assessment of an object's identity over time under varying parameters permits class predictions to be refined and improved iteratively.

LSTM networks fundamentally operate on temporally-based data series rendering them intrinsically oriented towards the goals of this task. In application to the evaluation of video and image sequences of length n, constituent individual image frames are considered sequentially. For some frame f_i, output from LSTM layer(s) are fed as input to layer(s) in the subsequent iteration for frame f_{i+1}. In the case of the task here, following processing of frame f_n, a final class-prediction vector is yielded for the entire input sequence via a fully connected layer.

Convolutional visual feature representations of input individual frames are extracted prior to input into a LSTM layer using the pool3-layer output of an Inception V3 CNN [56]. Such a combination of CNNs and LSTMs – first introduced by Donahue et al. [15] – is referred to as Long-term Recurrent Convolutional Networks (LRCNs) and is employed here for the task of spatio-temporal identification. Figure 4 illustrates the standard LRCN pipeline used. We propose that this core identification pipeline can then be easily integrated into a complete video processing architecture as illustrated in Figure 1, *(bottom)*.

5. Datasets

In order to train and evaluate the efficacy of proposed approaches, we make use of two datasets captured in practically relevant scenarios – namely *FriesianCattle2017* and *AerialCattle2017*. We publish both datasets for public use online at: http://data.bris.ac.uk.

5.1. FriesianCattle2017

This dataset comprises 940 RGB images based on data capture from the work preceding this paper [2]. Of those images, there are 89 distinct Holstein Friesian individuals. The data was captured over a two hour-long session via the use of a Microsoft Kinect 2 camera affixed statically over the walkway between holding pens and milking stations in a real indoor farming environment. The camera was configured to capture top-down still images of cattle dorsal coat patterns at a rate of 0.5 Hz. A more in-depth discussion of dataset acquisition, pre-processing steps, etc. is given in the original paper [2]. Example images of individuals from the dataset are given in Figure 5a.

5.1.1 Ground Truth Labelling

The use of R-CNNs intrinsically requires training data to include ground truth bounding boxes as well as class labels
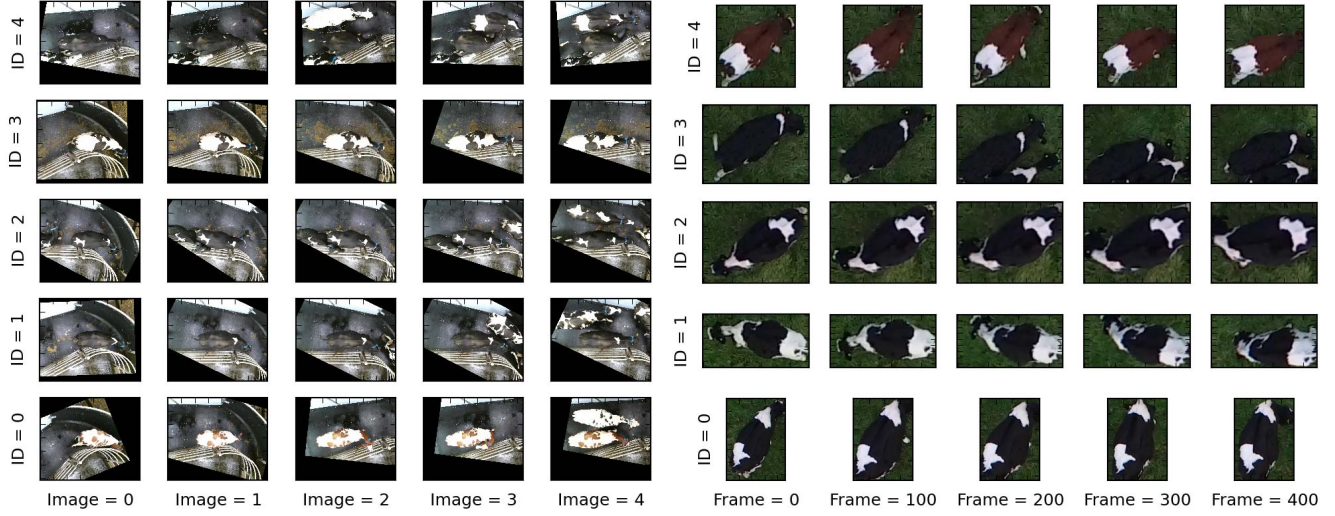
(a) **FriesianCattle2017 Dataset.** Examples of normalised still images. The camera was affixed statically above a walkway between holding pens and milking stations in a barn environment.

(b) **AerialCattle2017 Dataset.** Example frames from RoI streams extracted from outdoor video data and forwarded as input to the LRCN identification architecture.

Figure 5: **Dataset Examples.** Images from both indoor and outdoor datasets used in this work. *(rows)* Example individuals and *(columns)* example instances (or frames) for a particular individual.

for each object RoI. The dataset ground truth labelling process was therefore a two-fold task:

1. **Bounding Box Annotation**: After user annotation of bounding boxes in an image, the generated data was stored in an XML format aligned with that required by the Faster-RCNN framework for data annotation. Annotations were performed adhering to labelling guidelines of the VOC challenges [18] – specifically, the VOC2012 guideline [17].

2. **Individual/Class Labelling**: Following bounding box annotation, the users were sequentially presented each labelled RoI and asked to identify the individual cow contained within the presented bounding box.

5.1.2 Augmentation

During data capture, cows were free to walk from a holding pen to a milking station as they pleased. This results in per-individual variation in the time spent within view of the static acquisition system. Consequently, the number of images (or instances) obtained per individual are not balanced across the population – the mean being $\mu = 15$ images with standard deviation $\sigma = 19.9$.

To balance the number of images per individual for training purposes, the dataset was augmented via image synthesis. The target number of instances was chosen to be the maximum number of original (non-synthesised) instances for any particular individual (in this case, 127 images for Cow 11). Additional images were synthesised by rotating original images by some random angle α about the image centre (x_c, y_c) whilst maintaining the original image resolution for dataset consistency. Bounding box co-

ordinates were also transformed by angle α. However, since the Faster R-CNN implementation currently does not support the parameterisation of object rotation (bounding box angle), orthogonal bounding boxes were generated via min, max functions of transformed coordinates. The negative implication of this is that more background pixels are often included within ground truth object RoIs.

5.2. AerialCattle2017

The AerialCattle2017 dataset consists of 34 herd videos of cattle captured from an aerial standpoint of an outdoor agricultural field environment. Each video is approximately 20 seconds in length and is taken from a top-down perspective (see Fig. 2). For each video, cow regions were extracted and used to produce cropped videos containing single individuals (examples provided in Fig. 5b). Following individual cropping, the resulting regional dataset contains 23 individuals and a total of 160 videos for a mean of $\mu = 7$ instances per individual with standard deviation $\sigma = 3.87$.

5.2.1 Acquisition

This aerial video dataset was acquired via the use of a DJI Inspire MKI UAV or drone and its integrated camera/3-axis gimbal system. It was flown above a herd of approximately 30 young Holstein Friesians in a nursery field at the University of Bristol's veterinarian farm. During flights under the supervision of veterinarian researcher Dr Becky Whay, footage was captured over an hour-long period at a resolution of 3840x2160 pixels at 24 fps. Over the 34 acquired videos, the UAV height from the ground was varied by 5 m decrements starting at 25 m and with the lowest height
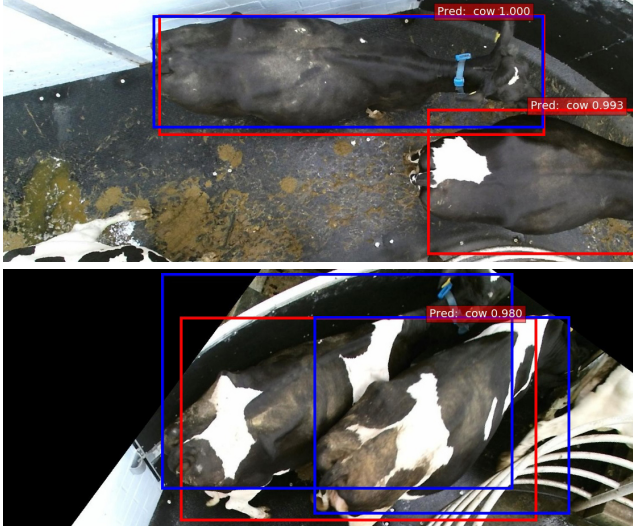
Figure 6: **Species Detection and Localisation Failures.** Shown are examples where cattle detection failed due to: *(top)* frame boundary clipping and *(bottom)* multi-cow alignment.

being 5 m – note that altitudes were varied only in-between video capture. This allowed the cattle to become accustomed and comfortable with the physical and sonic presence of the UAV – after initially exhibiting signs of some anxiety towards it.

5.2.2 Ground Truth Labelling and Cropping

To produce ground-truth bounding box and class labels, the first frame of each video was extracted and considered as for the indoor, still-image dataset. Object bounding boxes were manually annotated for these frames in accordance with labelling guidelines created for the VOC challenges [17, 18]. As for the aforementioned dataset, labelled bounding boxes were subsequently categorised into respective individual classes. To generate ground-truth bounding boxes for all subsequent frames of each video, an instance of the KCF tracking algorithm [27] was initialised for each RoI. Bounding box coordinates are updated over frames and were directly used to crop images to individual size. Clipped/lost regions were manually deleted to adhere to the VOC labelling guidelines. Cropped RoIs containing single cows were then saved as independent videos and grouped according to their original class label. Example frames yielded from this process are given in Figure 5b.

6. Results and Evaluation

In this section, results and corresponding evaluations are given for the three tasks of: (a) cattle/species detection and localisation, (b) single frame identification, and (c) video-based identification. All training and evaluation was conducted using a 3.6 GHz AMD 8-core Bulldozer CPU with an Nvidia GTX 1080Ti GPU and 32 GB of DDR3 RAM.

6.1. Cattle Detection and Localisation

We implement cattle detection and localisation using the VGG CNN M 1024 network adapted for the Faster R-CNN framework [49]. Written to support a Python API, the software implementation of Faster R-CNN is founded upon the Caffe deep learning library developed by Jia et al. [31].

The dataset used for evalutating detection and localisation performance of the R-CNN was formed as a union over frames from the FriesianCattle2017 and AerialCattle2017 datasets. This combination yields improved solution generalisation and avoids simplification of the detection task towards a single agricultural environment. All original (non-synthesised) images were used from the FriesianCattle2017 dataset. Every 12th frame (or a rate of 0.5 Hz) for each AerialCattle2017 video was extracted. This combination yielded the dataset consisting of 1077 images. We performed 2-fold cross validation over this set.

Region predictions from the R-CNN are accepted as true positives provided there is sufficient overlap with a same-class ground truth bounding box via a binary threshold $t_{ov} = 0.5$. Rectangle overlap is computed via Intersection-over-Union (IoU):

$$ov = \frac{bbox_{gt} \cap bbox_p}{bbox_{gt} \cup bbox_p} \tag{1}$$

where $bbox_{gt}$ and $bbox_p$ denote the ground truth and predicted bounding box regions respectively. Categorised detections are subsequently used to compute class precision and recall data. Mean Average Precision (mAP) is computed here via the Area under Curve (AuC) for generated precision-recall curves.

Table 1 summarises the 2-fold cross-validated performance tests, whilst Figure 7 illustrates a detail from the corresponding precision-recall curves. The evaluation demon-
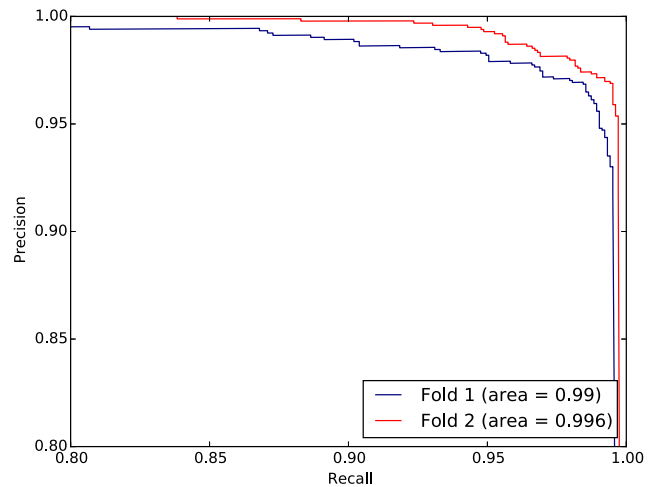

Figure 7: **Precision-Recall Curve for Detection.** Detailed section of the precision-recall curve for cattle detection and localisation for each of the two folds of cross validation.

strates that the task and data tested are well suited to the employed R-CNN framework. It produces near perfect results of correctly localising cows across the datasets. Figure 6 depicts some of the few examples where cattle detection failed. Examined failures consisted of: (a) erroneous detections created by the alignment and proximity of multiple cows or (b) false positive detections of partially-visible cattle having no corresponding ground-truth label due to the VOC labelling guidelines on object visibility/occlusion.

Task	mAP (%)		
	Fold 1	Fold 2	Average
Detection & Localisation	99.02	99.59	**99.3**

Table 1: **Species Detection Accuracy.** mAP values for 2-fold cross-validated performance tests for cattle detection and localisation. mAP scores are computed via AuC for generated precision-recall curves shown in Figure 7. The R-CNN cattle detector produces near perfect results for the task of correctly localising cows across the tested datasets.

6.2. Single Frame Individual Identification

For this task, the augmented (including synthesised images) FriesianCattle2017 dataset of indoor still images was used. Following synthesis on the 940 original instances
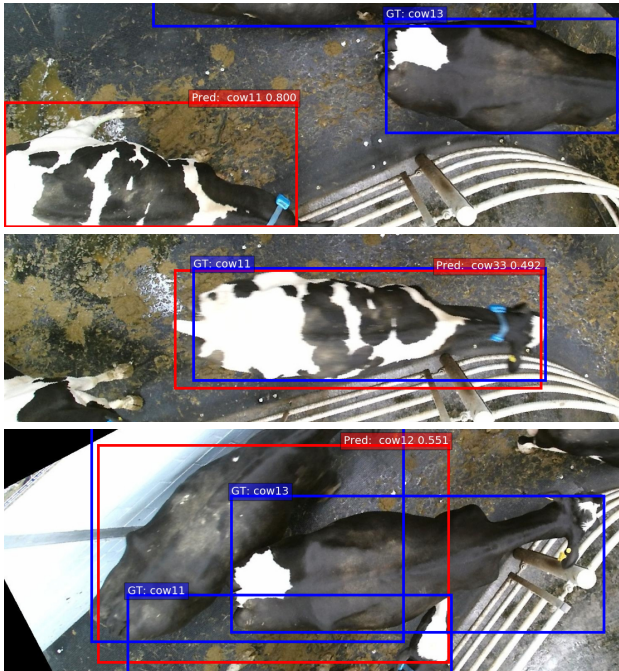

Figure 8: **1-Frame Identification Failures.** Shown are ground-truth/prediction examples where single-frame cattle identification failed due to: *(top)* no corresponding ground-truth annotation from following VOC labelling guidelines [17, 18], *(middle)* individual visual similarity and *(bottom)* multi-cow alignment/proximity.

for 89 individuals, the $\sim$11,000 yielded instances were randomly partitioned towards 2-fold cross validation. However, synthesised instances were never included in testing sets. Similarly to the previous task of cattle detection and localisation, the VGG CNN M 1024 network adapted for R-CNN was also employed. Training end-to-end for 100,000 iterations for both folds, the mAP values given in Table 2 were obtained. Figure 8 illustrates occasions where identification failed. These were found to be due to visual similarity across individuals and as for species detection; multiple cow proximity/alignment and frame boundary clipping.

Task	mAP (%)		
	Fold 1	Fold 2	Average
R-CNN Identification & Localisation	87.21	84.93	**86.07**

Table 2: **Single Frame Individual Identification Accuracy.** Classification mAP for single frame individual identification and localisation on the indoor FriesianCattle2017 dataset over 2-fold cross validation via the use of a R-CNN and the VGG CNN M 1024 network.

6.3. Video-based LRCN Identification

The dataset used for this task originates solely from the outdoor video dataset (AerialCattle2017) and consists of 46,430 cropped image frames and 23 individual cows over 160 total video instances. Video instances were split into 40-frame long spatio-temporal streams. Performing this stage for the entire data corpus resulted in 1064 labelled streams, each containing a single individual. This data was then partitioned[1] into a ratio of 9:1 for training and testing.

---

[1]Data partitioned using the `train_test_split` SciKit-Learn Python machine learning library function.
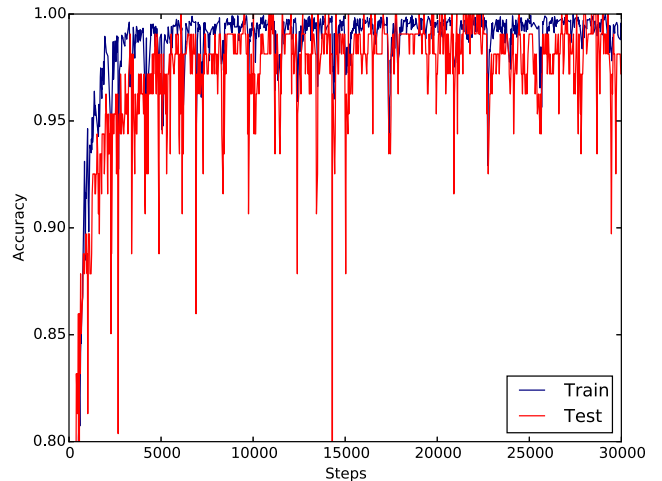

Figure 9: **Training towards Individual Identification.** Prediction accuracy for identification at different stages of LSTM training over 1,000 epochs on the training and testing datasets consisting of 957 and 107 streams, respectively.
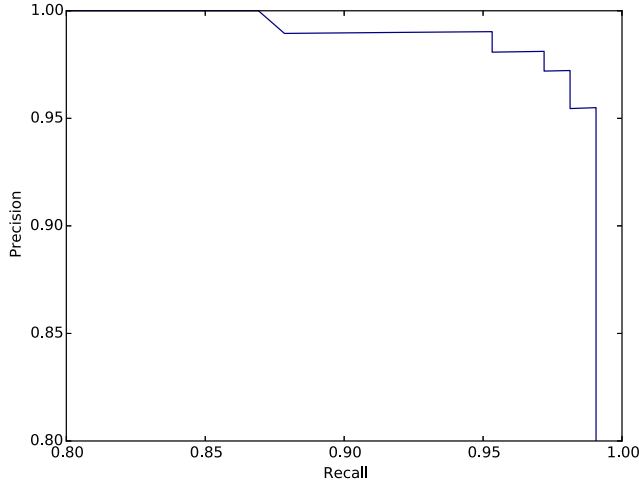
Figure 10: **Precision-Recall Curve for Video Identification Evaluation.** Detailed precision-recall curve for LRCN [15] identification on retained testing data consisting of 23 possible cow classes for 107 video streams.

To perform identification refinement, the Inception V3 network [56], an extension of GoogLeNet [55], was fine-tuned on the 23-class training dataset. In particular, each frame from the 957 training streams was used to fine-tune Inception originally trained for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [52]. Subsequently, all frames from each training stream were passed through the re-trained network. The 2048-wide vector yielded at the pool3-layer was captured in each case. Convolutional frame representations were then used to train a single LSTM layer comprised of 256 cells. Figure 9 illustrates prediction accuracies on the training and testing datasets over 1,000 epochs, with detailed identification results given in Table 3. For each prediction, an ordered vector of size $|classes| = 23$ is produced with class confidences $\in [0, 1]$. The predicted class label is taken to be the index of the maximum value in that vector. A prediction is then considered a true positive if the predicted and ground truth class labels match.

Figure 10 shows a detailed section of the precision-recall curve for the identification task, whilst some examples of false positive classifications are given in Figure 11. In the vast majority of failure cases, the individuals referred to by predicted and ground truth class labels were visually similar. The quantity (number of white, black or brown patches),

	Accuracy (%)	
Task	Training	Testing
LRCN Individual Identification	99.79	**98.13**

Table 3: **Video Identification Accuracy.** Classification accuracies of the LRCN setup after training for 1,000 epochs.

(a) Predicted and ground truth IDs = {3, 10} respectively.

(b) Predicted and ground truth IDs = {11, 14} respectively.

Figure 11: **Video Identification Failures.** Examples of false positive classifications from LRCN [15] identification on the AerialCattle2017 dataset. In all failure cases, dorsal features are very similar in structure and positional distribution on the body.

position and shape/structure of coat pattern features were found to bare strong resemblance across mistaken labels.

7. Conclusions and Future Work

This work demonstrates that the task of automatic Holstein Friesian detection and individual identification can be addressed via standard deep learning pipelines. For identification in particular, we show that convolutional-based architectures are well suited towards learning and distinguishing the properties of unique dorsal pattern and structure exhibited by the species individually. Importantly, this process can take place non-intrusively in practically relevant settings, in contrast to the majority of existing identification frameworks revolving around physical tagging.

Future work will include testing identification in fully independent captured scenarios to avoid possible data correlation completely. Throughout the identification process conducted here, the agent (UAV) was flown manually and merely captured data. It was not informed by the identification process and therefore can be seen as passive; captured video footage was analysed by the LRCN offline. Whilst we have shown that this can prove sufficient in the scenarios presented, future work will consider more complicated setups of faster moving and larger herds, as well as tight animal gatherings. In order to achieve sufficient identification performance in those scenarios and minimise animal disturbance, our work will begin to involve active robotic agents in an online identification process; that is, navigation will be directly informed by the progress of monitoring tasks and side parameters such as disturbance minimisation.

References

[1] A. Allen, B. Golden, M. Taylor, D. Patterson, D. Henriksen, and R. Skuce. Evaluation of retinal imaging technology for the biometric identification of bovine animals in Northern Ireland. *Livestock Science*, 116(1-3):42–52, 2008.

[2] W. Andrew, S. Hannuna, N. Campbell, and T. Burghardt. 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)*, pages 484–488, Sept 2016.

[3] A. C. Arslan and M. Akar. 3D Cow Identification in Cattle Farms. *Siu 2014*, (Siu), 2014.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-Up Robust Features. *Computer Vision and Image Understanding (CVIU)*, 110(September):346–359, 2008.

[5] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Trans. Neur. Netw.*, 5(2):157–166, Mar. 1994.

[6] M. B. Bowling, D. L. Pendell, D. L. Morris, Y. Yoon, K. Katoh, K. E. Belk, and G. C. Smith. Review: Identification and Traceability of Cattle in Selected Countries Outside of North America. *The Professional Animal Scientist*, 24:287–294, 2008.

[7] W. Buick. Animal passports and identification. *Defra Veterinary Journal*, 15:20–26, 2004.

[8] A. By. Breed differences and intra-breed genetic variability of dermatoglyphic pattern of cattle. 110:385–392, 1993.

[9] C. Cai and J. Li. Cattle face recognition using local binary pattern descriptor. *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4, 2013.

[10] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

[11] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014.

[12] J. Dai, Y. Li, K. He, and J. Sun. R-FCN: object detection via region-based fully convolutional networks. *CoRR*, abs/1605.06409, 2016.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[14] P. D'Hour, A. Hauwuy, J. Coulon, and J. Garel. Walking and dairy cattle performance. In *Annales de zootechnie*, volume 43, pages 369–378, 1994.

[15] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko. Long-term recurrent convolutional networks for visual recognition and description. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:2625–2634, 2015.

[16] D. S. Edwards, A. M. Johnston, and D. U. Pfeiffer. A comparison of commonly used ear tags on the ear damage of sheep. *Animal Welfare*, 10(2):141–151, 2001.

[17] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[19] R. H. Fayed and A. E. Hassanien. Muzzle Print Images. pages 529–534, 2013.

[20] F. Gers and J. Schmidhuber. Recurrent nets that time and count. *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, pages 189–194 vol.3, 2000.

[21] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[22] R. Girshick, J. Donahue, T. Darrell, U. C. Berkeley, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. pages 2–9, 2012.

[23] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[24] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[25] H. M. E. Hadad, H. A. Mahmoud, and F. A. Mousa. Bovines muzzle classification based on machine learning techniques. *Procedia Computer Science*, 65:864 – 871, 2015. International Conference on Communications, management, and Information technology (ICCMIT'2015).

[26] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv preprint arXiv:1406.4729*, 2014.

[27] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):583–596, 2015.

[28] S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. *Master's thesis, Institut fur Informatik, Technische Universitat, Munchen*, 1991.

[29] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.

[30] R. Houston. A computerised database system for bovine traceability. *Revue Scientifique et Technique*, 20:652–661, 2001.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, and U. C. B. Eecs. Convolutional_Architecture_Feature_Embedding. 2014.

[32] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. *Journal of Machine Learning Research*, 2015.

[33] N. Kalchbrenner, I. Danihelka, and A. Graves. Grid Long Short-Term Memory. pages 1–14, 2015.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[35] S. Kumar and S. K. Singh. Automatic identification of cattle using muzzle point pattern: a hybrid feature extraction and classification paradigm. *Multimedia Tools and Applications*, pages 1–30, 2016.

[36] S. Kumar, S. K. Singh, and A. K. Singh. Muzzle point pattern based techniques for individual cattle identification. *IET Image Processing*, (January), 2017.

[37] S. Kumar, S. K. Singh, R. S. Singh, A. K. Singh, and S. Tiwari. Real-time recognition of cattle using animal biometrics. *Journal of Real-Time Image Processing*, pages 1–22, 2016.

[38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[39] D. Lowe. Object Recognition fromLocal Scale-Invariant Features. *IEEE International Conference on Computer Vision*, 1999.

[40] C. Martinez-Ortiz, R. Everson, and T. Mottram. Video tracking of dairy cows for assessing mobility scores. *Joint European Conference on Precision Livestock Farming*, pages 154–162, 2013.

[41] P. V. Medicine, A. Adesiyun, and W. Indies. Ear-tag retention and identification methods for extensively managed water buffalo (Bubalus bubalis) in Trinidad for extensively managed water buffalo. (July):286–296, 2014.

[42] H. Minagawa, T. Fujimura, M. Ichiyanagi, and K. Tanaka. Identification of beef cattle by analyzing images of their muzzle patterns lifted on paper. *Publications of the Japanese Society of Agricultural Informatics*, 8:596–600, 2002.

[43] J.-M. Morel and G. Yu. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2:438–469, 2009.

[44] E. New, D. f. E. F. Zoonotic Disease (NEZD) Division, and R. Affairs. Most common breeds of cattle in gb (nuts 1 areas), 2005.

[45] A. Noviyanto and A. M. Arymurthy. Automatic Cattle Identification based on Muzzle Photo Using Speed-Up Robust Features Approach. *European Conference on Computer Science (ECCS)*, pages 110–114, 2012.

[46] A. Noviyanto and A. M. Arymurthy. Beef cattle identification based on muzzle pattern using a matching refinement technique in the SIFT method. *Computers and Electronics in Agriculture*, 99:77–84, 2013.

[47] E. Parliament and Council. Establishing a system for the identification and registration of bovine animals and regarding the labelling of beef and beef products and repealing council regulation (ec) no 820/97, 2000. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:32000R1760.

[48] W. Petersen. The Identification of the Bovine by Means of Nose-Prints. *Journal of Dairy Science*, 5(3):249–258, 1922.

[49] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *ArXiv 2015*, pages 1–10, 2015.

[50] A. Robinson, F. Fallside, and U. of Cambridge. Engineering Department. *The Utility Driven Dynamic Error Propagation Network*. University of Cambridge Department of Engineering, 1987.

[51] W. Rossing. Animal identification: introduction and history. *Computers and Electronics in Agriculture*, 24(1-2):1–4, 1999.

[52] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[53] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[54] C. Shanahan, B. Kernan, G. Ayalew, K. McDonnell, F. Butler, and S. Ward. A framework for beef traceability from farm to slaughter using global standards: An Irish perspective. *Computers and Electronics in Agriculture*, 66(1):62–69, 2009.

[55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[56] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[57] M. Tadesse and T. Dessie. Milk production performance of zebu, holstein friesian and their crosses in ethiopia. *Livestock Research for Rural Development*, 15(26), 2003.

[58] A. Tharwat, T. Gaber, A. E. Hassanien, H. A. Hassanien, and M. F. Tolba. Cattle identification using muzzle print images based on texture features approach. In *Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014*, volume 303 of *Advances in Intelligent Systems and Computing*, pages 217–227. 2014.

[59] J. F. Velez, A. Sanchez, J. Sanchez, and J. L. Esteban. Beef identification in industrial slaughterhouses using machine vision techniques. *Spanish Journal of Argricultral Research*, 11(4):945–957, 2013.

[60] D. D. Wardrope. Problems with the use of ear tags in cattle. *The Veterinary record*, 137(26):675, jan 1995.

[61] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339 – 356, 1988.

[62] K. Yao, T. Cohn, K. Vylomova, K. Duh, and C. Dyer. Depth-Gated LSTM. pages 1–5, 2015.