Computer Vision on the Edge: Individual Cattle Identification in Real-time with ReadMyCow System

Moniek Smink[1], Haotian Liu[1], Dörte Döpfer[2], Yong Jae Lee[1]
[1]Department of Computer Sciences, University of Wisconsin - Madison, United States
[2]School of Veterinary Medicine, University of Wisconsin - Madison, United States
smink2@wisc.edu, lht@cs.wisc.edu, dopfer@wisc.edu, yongjaelee@cs.wisc.edu

Abstract

In precision livestock farming, the individual identification of cattle is crucial to inform the decisions made to enhance animal welfare, health, and productivity. In literature, models exist that can read ear tags; however, they are not easily portable to real-world cattle production environments and make predictions mainly on still images. We propose a video-based cattle ear tag reading system, called ReadMyCow, which takes advantage of the temporal characteristics in videos to accurately detect, track, and read cattle ear tags at 25 FPS on edge devices. For each frame in a video, ReadMyCow functions in two steps. 1) Tag detection: a YOLOv5s Object Detection model and NVIDIA Deepstream Tracking Layer detect and track the tags present. 2) Tag reading: the novel WhenToRead module decides whether to read each tag, using a TRBA Scene Text Recognition model, or to use the reading from a previous frame. The system is implemented on an edge device, namely the NVIDIA Jetson AGX Orin or Xavier, making it portable to cattle production environments without external computational resources. To attain real-time speeds, ReadMyCow only reads the detected tag in the current frame if it thinks it will get a better reading when a decision metric is significantly improved in the current frame. Ideally, this means the best reading of a tag is found and stored throughout a tag's presence in the video, even when the tag becomes occluded or blurry. While testing the system at a real Midwestern dairy farm housing 9,000 cows, 96.1% of printed ear tags were accurately read by the ReadMyCow system, demonstrating its real-world commercial potential. ReadMyCow opens opportunities for informed data-driven decision-making processes on commercial cattle farms.

1. Introduction

Precision livestock farming (PLF), or the rearing of livestock informed by electronic sensors, tools, and methods, is spreading worldwide. PLF allows a producer to make early, optimal, and data-driven decisions informed by the monitoring of their livestock's behavior, welfare, and production [25]. These systems enhance animal welfare, health and productivity, as well as improve farming lifestyle, knowledge, and traceability of livestock products [26]. Furthermore, PLF has the potential to reduce the historically repetitive and physically demanding jobs conducted in adverse environments in the livestock sector [23], though further contemplation of the impacts of PLF is necessary [13]. Support of cattle health and welfare is particularly important as the cattle industry represents the largest share of total cash receipts for agricultural commodities in the U.S., forecasted to represent about 78.5 billion dollars in 2022 [8]. Thus, applications of AI are receiving increased attention in agriculture, particularly in smart agriculture, precision farming, and food animal health and production [40].

Many recent advancements have been made in the detection of cattle behavior and disease using AI. Models capable of recognizing heat [29], claw lesions [44], respiratory disease [41], mastitis [47], lameness [38], among others have found success in experimental settings. However, a gap exists between the development of disease detection models and their integration into real-time inference on commercial cattle farms, because, despite detection of disease, it remains a challenge to readily identify the individual diseased animal, particularly on large-scale cattle farms.

The current industry standard for individual cattle identification is radio frequency identification (RFID) using transponders. The radio signal of RFID tags is affected by electromagnetic environments such as iron fences and cooling fans as well as tag power and circuit quality [39]. According to [39], the average reading distance of low-frequency RFID tags is less than 8 inches, limiting the application of this identification method, because disease detection models and strategies are applied at larger distances. In addition, ear tags with visual identification numbers are cheaper, limiting the widespread use of RFID tags. An alternative individual cattle identification system is needed.

Visual cattle identification tags, applied as ear tags or neck bands, use printed or handwritten sequences of num-

bers to identify cattle and other farm animals. The number of digits on the tags, tag warping, their color, lighting, filth, and cattle movement make reading these ear tags using optical character recognition (OCR) algorithms difficult. Models capable of individual identification of cattle exist and are functional in certain use cases [1,2,16,17,34,48,51,53,54], however, they are image-based and not portable.

Video-Based Recognition versus Image-Based Recognition: For the purposes of this paper, an image-based pipeline is an AI pipeline that makes predictions for a single frame using only information from that frame. A video-based pipeline is an AI pipeline that makes predictions for a frame while taking advantage of the temporal characteristics present in a video: it uses the information from past frames to make better predictions for the current frame. A single frame from a video can only provide so much information to an identification system. For example, if the last digit of an ear tag is occluded in a frame, an image-based tag reading pipeline will report the wrong reading for that frame, regardless of whether the correct reading had already been found in a previous frame. A video-based pipeline, on the other hand, can read this tag in its current state, compare this frame's reading to previous readings of the same tag through a tracker, and choose which reading to report based on confidences or other metrics. In addition, a video-based pipeline is capable of skipping certain computationally expensive steps in certain frames where they are deemed unnecessary, speeding up the pipeline. The cattle identification models currently present in literature are mostly image-based.

Portability-Performance Tradeoff: A limitation of current cattle identification models is their portability. Computer vision models often demand extensive computational resources. Most commercial cattle farms do not have access to these computational resources or a stable network connection. A solution to this problem is to run AI pipelines on edge devices. However, the smaller and more convenient the device, usually the more limited its computational power, slowing detection speed and reducing the model's effectiveness, although, with modern advances such as TinyML [50], this tradeoff could diminish. NVIDIA has started integrating their high-powered GPUs into a line of edge devices capable of efficiently running accurate AI models. An effective and useful cattle identification model must strike a balance between portability and performance, where environmental constraints and model effectiveness coexist.

Contributions: We propose a tag reading system, called ReadMyCow, capable of detecting and tracking multiple visual ear tags throughout their consecutive appearances in a video stream, reporting the best tag reading, at 25 FPS on an edge device. A key technical contribution is the WhenToRead module, which decides whether a tag should be read anew in each frame or to use a previous frame's reading, enabling real-time speeds. Our experiments demonstrate that our system is significantly faster and more accurate than both an image-based baseline as well as a video-based baseline that only performs tracking without the WhenToRead module. The ReadMyCow system provides an important step towards mobilizing AI models in applied settings as well as introducing a tag reading system capable of commercial utilization in a cattle production setting, enabling exact animal husbandry in precision livestock farming. When presented and tested at a commercial 9,000 cow dairy farm, the management of the dairy expressed interest in the ReadMyCow system, asking how soon the system would be available commercially. Importantly, while we focus on cows due to the data that we have, our approach could also easily be applied to other livestock animals such as pigs and sheep.

2. Related Work

2.1. Individual Identification of Cattle

Physical Body Features: Reports in literature propose individual cattle identification through body markings [3, 16, 33, 34, 51]. Similarly, [1, 48] use bovine faces for identification. These identification methods are limited to the cattle present at the time of training the model, reducing the scalability of these models on commercial farms with large turn-over of cattle. [2] addresses this by proposing a more open and general model capable of differentiating cows it has never seen before. Yet, mapping the differentiated body markings to an individual animal is still necessary. In general, when faced with a commercial production environment where the number of cows can exceed 1000, the scalability and usefulness of cattle identification models that use body features are limited. In addition to cattle, it is important to mention that a lot of work has been done in the individual identification of other animals such as sheep [45], pigs [14,28], and wild animals using a wide variety of methods [20, 43].

Ear Tags: Most large-scale cattle production farms already use individual identification numbers. Animals wear ear tags with a visible unique ID that is familiar to the farmer. As RFID tags are more expensive and cannot be read from a distance, researchers have proposed visual ear tag identification systems. [17] proposes a system that can read one near ear tag at a time using color thresholds, flood fill, Hough transform, skewness correction, and projection methods. [53, 54] propose an ear tag reading system that functions by the feed bunks, capable of simultaneously reading 4-digit ear tags of five cows in pre-determined locations using head detection, HSV conversion, color thresholds, skewness correction, digit segmentation, and digit

recognition using a Convolutional Neural Network (CNN). [6] set up a model in a milking robot, reading ear tags by zeroing in on the tag through HSV conversion, color thresholding, and skewness correction, followed by acquiring the readings using a CNN fine-tuned on tag data.

These individual cattle identification models are functional but limited. 1) They are trained for very specific conditions and not sufficiently flexible to adapt to a real farm environment. For example, away from the feed bunk or milking robot, when the colors of the tags are not uniform, when the number of digits on the tags vary, or in situations where the tag is obscured through motion, occlusion, or dirt. 2) They suffer from the image-based versus video-based recognition problem: the recognition pipeline is restricted to operating on a frame-by-frame basis, without tracking the readings of the same tag from previous frames in the video stream. This means that in frames where a correct reading is simply impossible, for example, due to a pole blocking a digit, blur from a flicking ear, among others, the system will report a wrong reading, even if the correct reading had already been given a few frames earlier. 3) Another limitation of these tag reading models is the portability-performance tradeoff: the models are decently fast, the detections are mostly accurate, but the models are not portable to a production cattle setting where computational resources and internet connections are limited.

2.2. Deep Learning Approaches to Object Detection

Object detection research has had remarkable advances over the past decade. Notable deep learning methods include those of the R-CNN [11, 37] and YOLO [21, 35, 36] families. Video object detection methods combine static image-based object detection with video-based tracking to leverage the temporal signal in video for improved accuracy and/or speed [10, 19, 22, 46, 52]. Our work builds upon the YOLOv5 detector for its high accuracy and real-time speeds, and the NVIDIA Deepstream Tracking Layer for tracking. Importantly, we introduce a novel WhenToRead module, which exploits temporal redundancy in video to further improve accuracy and speed, and attain real-time speeds on edge devices.

2.3. Deep Learning Approaches to Optical Character Recognition and Scene Text Recognition

Optical character recognition (OCR) has been studied for decades [7, 49]. Scene text recognition (STR) is a specific type of OCR problem in which the task is to recognize text (often whole words instead of individual characters) in natural images [24]. The deep learning era, combined with popular competitions such as the ICDAR Robust Reading Competition [30], have led to significant progress in STR both on the accuracy and inference speed fronts. STR methods have traditionally been trained with human-annotated real images. To reduce the dependency on expensive hand-annotated data, researchers have explored using synthetic datasets [12, 18] as well as semi-supervised learning techniques [5, 9]. Existing approaches typically train a sequence prediction model like an LSTM [15] or Transformers [42]. In this paper, we leverage the TPS-ResNet-BiLSTM-Attn model [4] for its state-of-the-art accuracy and speed.

3. Methodology

Our ReadMyCow system takes in an input video of cows with eartags, and outputs bounding box detections of the eartags and their corresponding text readings. The ear tags can be on either ear or both, and the text can be either handwritten or printed.

The ReadMyCow system is comprised of two steps. For each frame in a video stream: 1) Tag detection: cattle ear tags are localized with bounding boxes and connected to tags from previous frames through a tracking number. 2) Tag reading: using the novel WhenToRead module, a decision is made whether each tracked bounding box should be read anew during this frame, or whether a previous reading should be used, increasing the accuracy and speed of the system by sometimes skipping the tag reading step in frames where reading anew may be unnecessary or provide an inaccurate reading. The overall system is implemented on an NVIDIA Jetson edge device. We provide further details on each step below. Fig. 1 shows the system overview.

3.1. Tag Detection

3.1.1 Preliminary Detection

The preliminary detection and localization of near-enough-to-read tags is done by a fine-tuned YOLOv5s model [21]. YOLO (*you only look once*) models are a popular line of object detection and image classification models known for their speed and accuracy. These models process an entire image in a single forward pass of a convolutional neural network (CNN), making them very fast. The YOLOv5 model appears in different sizes. For the purposes of the ReadMyCow system, we choose the 'small' size as a compromise between speed and accuracy. An advantage of leveraging the YOLOv5 line of models is that they are widely used in both academia and industry, and can be implemented within the NVIDIA Deepstream library—which we use for our tracker—on NVIDIA edge devices at very high speeds.

To adapt the pre-trained YOLOv5s object detection model to cattle ear tag detection, we first fine-tune the baseline model on a custom ear tag dataset. The custom ear tag dataset consists of 1,453 images of beef and dairy cattle with ear tags of different colors in varied situations on two farms housing more than 30,000 cattle total. Ear tags are labelled as either 'near' or 'far.' A 'near' tag is an ear tag close enough to the camera to be readable by the investigator; all

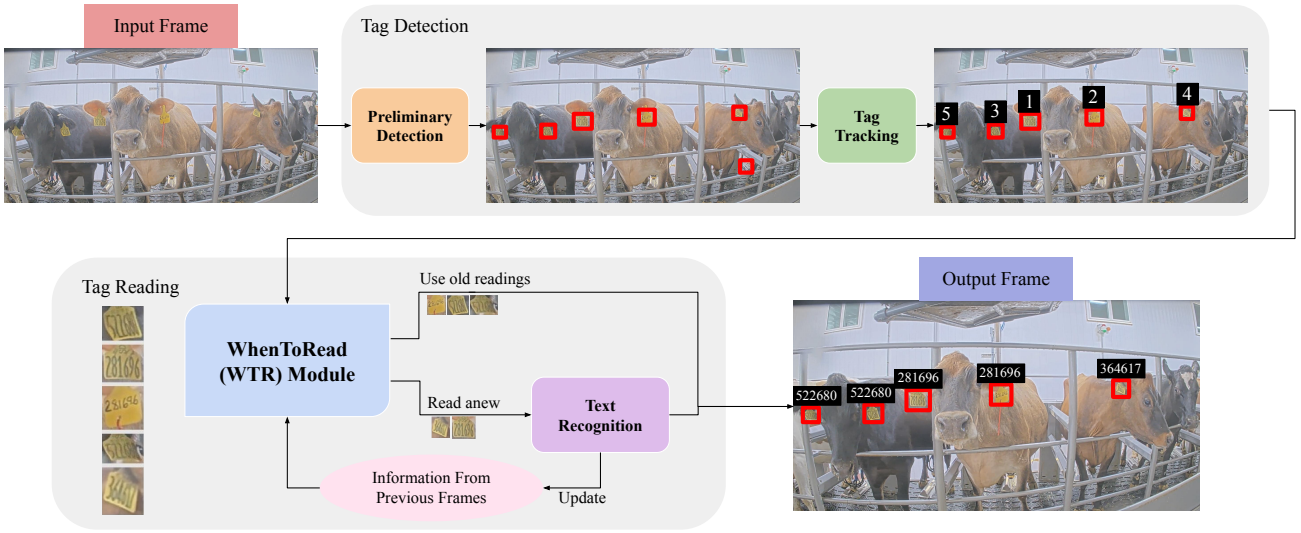

Figure 1. ReadMyCow system operations pipeline for a single frame in a video stream.

other tags are labeled as 'far.' Challenges present while labeling this dataset include defining the edge case procedure. Specifically, what to do with severely occluded and nearly illegible tags, and the sheer quantity of possible data; every farm has countless potential camera locations. At the request of the livestock farms, public access to this dataset is currently not possible, but an abbreviated dataset can be made available to individuals upon request. We randomly split the dataset into training (1,308 images) and test (145 images) sets, and fine-tune the baseline YOLOv5s model for 100 epochs. We keep the model with the best accuracy. For the best model, the mean average precision (mAP) of the 'near' and 'far' tags is 0.937 and 0.676, respectively.

3.1.2 Tracking of Detections

After a video frame's tags are detected by the YOLOv5s model, the extracted tags enter a multi-object tracking layer created by NVIDIA in their Deepstream library [31]. The main objective of this layer is to assign the same tag the same tracking identification number throughout the video stream. It does this by tracking each box as it appears throughout successive frames, while also suppressing potential false detections. This layer evaluates each detection bounding box for the likelihood of it being a false positive. A false positive detection would be a bounding box that does not actually contain a tag. If the tracking layer has little confidence in the detection truly being a tag, the detection can be suppressed until the confidence rises. A suppressed detection does not get sent to the tag reading step or get shown to the user. Once the tracking layer is confident that the tag detection is a true positive, a tracking number is assigned, and the tag enters the WhenToRead module in the tag reading step.

The tracking layer described above enables the ReadMyCow system to take advantage of video temporality, speeding up the system while also increasing its prediction accuracy. Low-confidence detections are suppressed, meaning they are not sent to the computationally expensive tag reading step, decreasing the amount of time spent on potentially unimportant detections. High-confidence detections are tracked, meaning the output readings from the tag reading step can be saved for each tracking ID, enabling the system to skip reading the same tag again when the reading is not likely to be improved in terms of accuracy. We evaluate the performance of the tag tracking step in Section 4.

3.2. Tag Reading

3.2.1 WhenToRead Module

Individual video frames may suffer from occlusion, motion blur, and corruption, among others. To ensure better tag recognition quality, we design a novel WhenToRead (WTR) Module, which dynamically selects the best frame for tag reading as the input stream continues. We show the design of the WTR module in Figure 2.

For each tracked tag detection in the input frame, the WTR module decides whether to read the tag anew or use an existing reading through a custom decision metric (DM). Specifically, we use the tag detection confidence scaled with the bounding box surface area as the DM. We choose this value as the DM because it quantifies when a tag could be more readable. As the surface area of the bounding box increases, the tag could be getting closer to the camera, rotating from diagonally facing forward to entirely facing for-

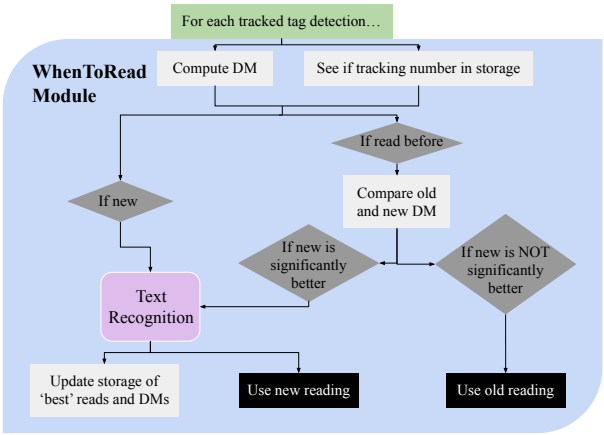

Figure 2. The internal procedure of the WhenToRead (WTR) module for a single tag detection, deciding whether the tag should be read anew or whether the old reading should be used based on a custom decision metric (DM).

ward, or becoming less occluded. The tag detection confidence can signal the absence of blur, motion, or warping. When a new tracked detection comes, we always read its content, compute the DM, and save these values. Subsequent readings of the same tag occur only if the current frame's DM significantly exceeds the previous 'best' reading's DM. For a DM to significantly exceed another, it must be larger than the other scaled by a sensitivity multiplier. A higher multiplier results in less frequent tag readings. In our implementation, we set the multiplier to 1.1 based on qualitative observations. Essentially, a tag is read again when it gets significantly closer to the camera and the system is more confident in its detection.

The WhenToRead module helps to increase and balance the speed and accuracy of the system. The system only reads the tags during frames predicted to generate a better reading, reducing the average computing time per frame. Additionally, in frames where the tag is less likely to yield a better reading (like when the tag is being slowly turned or being occluded), the tag will not be read again. Thus, ideally, the best reading for a tag is found and kept throughout that tracked tag's lifecycle in the video.

3.2.2 Text Recognition

Text recognition of tags is done by a fine-tuned Scene Text Recognition model called TPS-ResNet-BiLSTM-Attn (TRBA), as described in [4]. This model consists of a thin-plate spine (TPS) input image normalizer to straighten out text, a ResNet feature extractor, BiLSTM sequence modeler, and attention-based sequence predictor (Attn) [4]. We choose this model for the ReadMyCow system because it has the best accuracy in [4]. We choose a Scene Text Recognition (STR) model as the tag reader instead of an Optical Character Recognition (OCR) model because we want the system to function no matter how many digits are present on the ear tag.

We fine-tune the TRBA model on the custom ear tag dataset. From the 1,453 labelled cow images in the training dataset used to train the YOLOv5s model, the near-enough-to-be-read tags are cropped out and labelled with the correct tag readings. This results in 1,581 labelled images of tags with anywhere from two to six digits per tag. We randomly split this dataset into training (1,424 images) and test (157 images) sets. We train the TRBA model on the training set for 3,000 iterations, and choose the TRBA model with the best accuracy on the test set (75.16%).

3.3. Deployment on Edge Devices

We implement our system on a Jetson AGX Xavier and a Jetson AGX Orin. The Xavier is flashed with Jetpack version 5.1.1 while the Orin is flashed with Jetpack version 5.0.1. We install NVIDIA Deepstream SDK version 6.1 on both edge devices [31]. In order to use the YOLOv5s model within Deepstream, we leverage the Deepstream-Yolo library [27]. In order to write the software in Python, we use the deepstream python app library [32]. RTSP cameras (security cameras), video files, or USB cameras are used as video stream inputs.

4. Experiments

In this section, we evaluate the impact of the WhenToRead module on the accuracy and speed of the ReadMyCow system by comparing it to variations of the system without the module in an ablation study. We also evaluate the overall applicability of the ReadMyCow system by measuring what proportion of cows and tags it can identify when implemented on a commercial dairy farm. Finally, the usability of the ReadMyCow system in real-world cattle production is commented on by briefly sharing the ideas, contexts, and challenges of a commercial dairy farm.

4.1. Implementation Details

We set up the ReadMyCow system on a Midwestern commercial dairy farm with 9,000 Jersey and Holstein Fresian crossbreed cows. Cows are milked twice per day in a large rotating milking parlor with 106 stalls. A metal bracket holding a security RTSP camera is installed on the inside of the rotating parlor. While cows are milked, they stand facing the camera, rotating in and out of frame about every four seconds. Between zero and five cows could be found in frame at a time. We placed the ReadMyCow system, implemented on a NVIDIA Jetson AGX Orin, in a separate room about 20 meters away from the security camera. The RTSP camera streams into the ReadMyCow system. Further experimental details are described below.

4.2. Evaluating the WhenToRead Module

We first perform ablation studies to evaluate the impact of the WhenToRead (WTR) module on the ReadMyCow system. Specifically, three variations of the ReadMyCow system are implemented on a NVIDIA Jetson AGX Xavier: 1) No Tracking – No WhenToRead (No T, No WTR): for every frame, detected near-enough-to-read (NETR) tag bounding boxes are sent directly to the tag reading model to be read (image-based system without memory of previous frames – no false positive limiter). 2) Tracking without WhenToRead (T, No WTR): for every frame, detected NETR tag bounding boxes are sent to the tracking layer where potential false positives are suppressed; boxes with higher detection confidences are assigned a tracking ID and sent to tag reading model to be read (image-based system without memory of previous frames with false positive limiter). 3) Tracking with WhenToRead (T, WTR): for every frame, detected NETR tag bounding boxes are sent to the tracking layer where potential false positives are suppressed; boxes with higher detection confidences are assigned a tracking ID, and based on this tracking ID the WhenToRead module determines whether to read this tag or use a previous reading (full video-based ReadMyCow system as described in the Methodology section).

To compare these variations, we recorded a one-hour long video of 550 cows passing the RTSP camera in the rotating milk parlor. This one-hour video is the basis for the following evaluations: 1) Speed: comparing the FPS of the different modules; 2) Accuracy: comparing the tag reading accuracy of the different variations. See Fig. 3 for an overview of the results of these evaluations. Further details on each specific evaluation can be found below the figure.

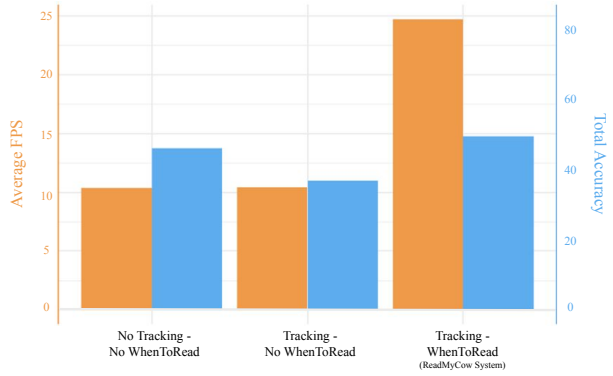

Figure 3. Comparing the speed and accuracy of three variations of the ReadMyCow system in an ablation study, demonstrating the impact of the WhenToRead (WTR) module and tracking layer. The average FPS is computed from a one-hour evaluation video of cows in a rotating milk parlor. The total text recognition accuracy is computed from 600 random test frames from the previously mentioned evaluation video.

1. Speed: each variation of the ReadMyCow system is applied to the one hour evaluation video. Every 5 seconds, we evaluate and record the FPS (frames per second). The average FPS over the length of the video and 2.5th/97.5th FPS quantiles can be found in Table 1.

System Variation	Average FPS	2.5th Quantile	97.5th Quantile
No T, No WTR	10.34	7.59	16.78
T, No WTR	10.40	7.40	20.62
T, WTR	**24.69**	**20.05**	**29.17**

Table 1. An ablation study: three variations of the ReadMyCow system applied to a one-hour video of cows being milked, demonstrating the impact of the WhenToRead (WTR) module and tracking layer (T) on the system speed. The FPS is computed every five seconds. Higher is better.

2. Accuracy: 600 frames from the one hour evaluation video are randomly selected. We apply each variation of the ReadMyCow system to the video, and extract the predictions for the 600 frames. We define a 'readable tag' as an ear tag that is readable by a human in any frame in the video, but not necessarily from one of the 600 test frames. For example, even if a test frame tag's last digit is occluded, it is still considered 'readable' if the correct reading can be found in a different frame. For each variation of the system, the number of readable tags found and the percentage of readable tags read successfully are shown in Table 2.

System Variation	% Tags Found	Found Tag Accuracy	All Tag Accuracy
No T, No WTR	**99.5**	46.1	45.8
T, No WTR	76.0	48.3	36.7
T, WTR	76.0	**64.6**	**49.1**

Table 2. An ablation study: three variations of the ReadMyCow system reading 613 'readable tags' from 600 test frames randomly extracted from a one-hour video of cows milked in a rotating parlor, demonstrating the impact of the WhenToRead (WTR) module and tracking layer (T) on the system's prediction accuracy. A 'readable tag' is defined as an ear tag that is readable to a human investigator in any frame of the video.

The WhenToRead module is able to more than double the average speed of the ReadMyCow system without losing its accuracy, even slightly improving it. A video-based recognition system that remembers the best reading from previous frames is able to be accurate even when a tag is occluded in a current frame, while an image-based approach cannot accomplish this.

We found that a total of 23 false positive tag detections are present in the No Tracking – No WhenToRead variation, which are not present in the other variations. This

shows that the tracking layer does indeed limit the presence of false positives. However, it also is sometimes too restrictive, suppressing readable ear tags and causing them to go unread. This gap of missing readable tags can be seen by comparing the accuracy of the No Tracking – No WhenToRead variation and the Tracking – No WhenToRead variation; the presence of the tracking layer actually decreases the accuracy of the model. The presence of the WhenToRead module more than makes up for this deficit, even slightly increasing the accuracy over the No Tracking – No WhenToRead variation.

4.3. Real World Use of ReadMyCow System

We next evaluate the ReadMyCow system's usability in a real-world setting. Specifically, we set up and run the system in a rotary milking parlor on a commercial Midwestern farm. We use one hour of real-time recognition. Fig. 4 depicts some example frames with real system outputs.

4.3.1 Quantitative Results

We report accuracies for two types of entities: tags and cows.

1. Tags: In the one-hour video, a total of 646 tags are humanly readable. We define a tag as 'humanly readable' if a human investigator playing back the video can pause, digitally zoom in, and read the corresponding tag in at least one frame. 461 of the tags are printed (71.4%); 185 tags are handwritten (28.6%). The overall proportion of tags that are read correctly is 84.2%. The overall proportion of printed tags that are read correctly is 96.1%. Handwritten tags are much more difficult to read than the printed tags with an overall accuracy of 54.6%. Results are shown in Table 3.

Type of Tag	Count	Accuracy
Printed	461	96.1
Handwritten	185	54.6
Total	646	84.2

Table 3. ReadMyCow system identification accuracy for 646 cattle ear tags in a one hour video of a rotating milk parlor.

2. Cows: Each cow has one, two, or zero ear tags, and can have any mixture of handwritten and printed ear tags. In the one-hour video, a total of 550 cows pass by the RTSP camera. We define an identifiable cow as a cow who has an ear tag that is humanly readable. 435 of the cows are identifiable (79.1%). Of the 115 unidentifiable cows, 86 never show the flat face of their ear tags, and 29 of them show their ear tags, but the tag number is unreadable by a human. Of all the 550 cows, 71.3% are correctly identified by the ReadMyCow system. Of the 435 identifiable cows, 90.1% are correctly identified by the ReadMyCow system. Of the 333 cows identifiable by a printed ear tag, 95.8% are correctly identified by the ReadMyCow system. The handwritten tags are again more difficult to read: of the 119 cows identifiable by a handwritten ear tag, 75.6% are correctly identified by the ReadMyCow system. Table 4 shows these results in detail.

Type of Cow	Count	Accuracy
Identifiable through print	333	95.8
Identifiable through handwriting	119	75.6
Identifiable	435	90.1
Total	550	71.3

Table 4. ReadMyCow system identification accuracy for 550 cows in a one hour video of a rotating milk parlor. An 'identifiable' cow is defined as a cow whose tag number can be read by a human investigator in any frame of the video.

4.3.2 Qualitative Results

The ReadMyCow system works extremely well on the good quality tags that are visible. However, several factors can still limit its accuracy:

1. Cow behavior: cows rarely put their ears forward, unless stimulus piques their interest. Some cows never show their ear tags to the camera at all, while others only flick their ears forward for a frame or two before flicking back. An idea for improving this limitation would be to tie or mount something interesting in front of the cows; however, this would be unsustainable in the long term when the cows lose interest. Another idea for mitigating this limitation would be to have another camera situated on the side of the cows to catch the sideways ear tags. Further testing would need to be done on evaluating the system's performance using multiple cameras.

2. Quality of the ear tags: many ear tags are handwritten. The model often has issues discerning 2s from 7s and 4s from 9s on handwritten tags. The model also has trouble adjusting to ear tags with uncentered numbers. Other ear tags are notched by herdsmen to signal treatments of diseases. Others are very bent or warped, resulting in half-hidden tags even when the cow points its ears forward. Some tags are blank. Fig. 5 shows examples of ear tags that were difficult for the system to read. An idea for improving some aspects of this limitation would be to further fine-tune the tag reading model on a bigger dataset including handwritten tags, damaged tags, and tags where the digits are not located in the center.

Figure 4. Example output frames from ReadMyCow system operating in real dairy farm. The bounding box line thickness and size of the text predictions have been enlarged for visibility.

Figure 5. Examples of difficult to read cattle ear tags.

Overall, the system felt dependable. If a tag is reasonably written and shown for a few frames, the system can accurately read it in real-time speed.

4.4. Usability in Real-World Cattle Production

To our pleasant surprise, the staff members of the commercial dairy farm were intrigued and excited about the models. The ReadMyCow system could be used in several locations on the farm: 1) By a series of gates to automatically direct identified cows to the correct pen. 2) In the maternity pen, when combined with another AI model, to identify calving cows. 3) In the rotating milk parlor, when combined with another AI model, to detect cow foot diseases. 4) In the general housing pens, to generate treatment lists for cows diagnosed by other disease detection systems.

The commercial dairy farm has to invest a lot of time and human resources into these repetitive tasks. For example, every 20 minutes, a staff member walks along the maternity pen to look for calving cows. Automating these tasks would be of great help to the farm and the cattle. The commercial dairy farm was impressed by the system's ability to accurately read many tags at once from a distance and looked forward to seeing the future developments of the ReadMyCow system.

Implementation Challenges: Every livestock farm is different: cattle breeds, lighting, tag colors, among others can all vary widely, as well as the place the farmer wants to implement the system. For the ReadMyCow system to be more broadly effective, much larger datasets of ear tags from widespread cattle farms should be created. Broadening the system to use multiple cameras should also be explored to increase the likelihood of successful detection. For each farm, local experts must establish ground rules for when the system is unable to identify a cow depending on the intent behind the system implementation. For example, if the goal is to send specific cows to specific pens as they return from the milking parlor, which gate should be opened for an unidentifiable cow? There is a long road to widespread implementation of the ReadMyCow system, but the system's flexibility highlights the potential for its application to many settings, tasks, and animals.

5. Conclusion

The ReadMyCow system can accurately identify cattle in real-time in a real-world environment. The WhenToRead module strikes a balance between accuracy and speed while also taking advantage of the temporality of videos to allow these computationally expensive models to run efficiently on edge devices. The ReadMyCow system does video-based recognition: meaning that the information from previous frames of a video can be used to inform decisions made in future frames. Consequently, it is able to identify a cow, even if its tag is occluded in the current frame. The ReadMyCow system is implemented on an edge device, meaning it is portable to farm environments without prior computational resources, and reads tags at 25 FPS.

When presented at a commercial dairy farm with 9,000 cows, the ReadMyCow system garnered interest from the management of the dairy, creating the potential to be permanently implemented. The ReadMyCow system opens opportunities for informed data-driven decision-making on commercial cattle farms, enhancing animal welfare, health, and productivity, as well as improving traceability of livestock products, knowledge, and farming lifestyle.

References

[1] Brahim Achour, Malika Belkadi, Idir Filali, Mourad Laghrouche, and Mourad Lahdir. Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on convolutional neural networks (cnn). *Biosystems Engineering*, 198:31–49, 2020. 2

[2] William Andrew, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. Visual identification of individual holstein-friesian cattle via deep metric learning. *Computers and Electronics in Agriculture*, 185:106133, 2021. 2

[3] William Andrew, Colin Greatwood, and Tilo Burghardt. Visual localisation and individual identification of holstein friesian cattle via deep learning. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 2850–2859, 2017. 2

[4] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. *CoRR*, abs/1904.01906, 2019. 3, 5

[5] Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *CVPR*, 2021. 3

[6] John W.M. Bastiaansen, Ina Hulsegge, Dirkjan Schokker, Esther D. Ellen, Bert Klandermans, Marjaneh Taghavi, and Claudia Kamphuis. Continuous real-time cow identification by reading ear tags from live-stream video. *Frontiers in Animal Science*, 3, 2022. 3

[7] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang. Text recognition in the wild: A survey. *ACM Computing Surveys*, 54(2), 2021. 3

[8] USDA ERS. Sector at a glance, Sep 2022. 1

[9] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *CVPR*, 2021. 3

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017. 3

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 3

[12] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016. 3

[13] Oleksiy Guzhva, Janice M. Siegford, and Christina Lunner Kolstrup. The hitchhiker's guide to integration of social and ethical awareness in precision livestock farming research. *Frontiers in Animal Science*, 2, 2021. 1

[14] Mark F. Hansen, Melvyn L. Smith, Lyndon N. Smith, Michael G. Salter, Emma M. Baxter, Marianne Farish, and Bruce Grieve. Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98:145–152, 2018. 2

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8), 1997. 3

[16] Hengqi Hu, Baisheng Dai, Weizheng Shen, Xiaoli Wei, Jian Sun, Runze Li, and Yonggen Zhang. Cow identification based on fusion of deep parts features. *Biosystems Engineering*, 192:245–256, 2020. 2

[17] Maja Ilestrand. *Automatic Eartag Recognition on Dairy Cows in Real Barn Environment*. PhD thesis, Linköping University, 2017. 2

[18] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1), 2016. 3

[19] Licheng Jiao, Ruohan Zhang, Fang Liu, Shuyuan Yang, Biao Hou, Lingling Li, and Xu Tang. New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8), 2022. 3

[20] Thani Jintasuttisak, Andrew Leonce, Moayyed Sher Shah, Tamer Khafaga, Greg Simkins, and Eran Edirisinghe. Deep learning based animal detection and tracking in drone video footage. In *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*, ICCAI '22, page 425–431, New York, NY, USA, 2022. Association for Computing Machinery. 2

[21] Glenn Jocher. YOLOv5 by Ultralytics, May 2020. 3

[22] Kai Kang, Hongsheng Li, Tong Xiao, Wanli Ouyang, Junjie Yan, Xihui Liu, and Xiaogang Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. 3

[23] K Sakthiaseelan Kumaraveloo and Christina Lunner Kolstrup. Agriculture and musculoskeletal disorders in low- and middle-income countries. *Journal of Agromedicine*, 23(3):227–248, 2018. PMID: 30047854. 1

[24] Shangbang Long, Xin He, and Cong Yao. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129, 2021. 3

[25] Daniela Lovarelli, Jacopo Bacenetti, and Marcella Guarino. A review on dairy cattle farming: Is precision livestock farming the compromise for an environmental, economic and social sustainable production? *Journal of Cleaner Production*, 262:121409, 2020. 1

[26] Daniela Lovarelli, Lorenzo Leso, Marco Bonfanti, Simona Maria Carmela Porto, Matteo Barbari, and Marcella Guarino. Climate change and socio-economic assessment of plf in dairy farms: Three case studies. *Science of The Total Environment*, 882:163639, 2023. 1

[27] Marcos Luciano. Deepstream-Yolo. 5

[28] Mathieu Marsot, Jiangqiang Mei, Xiaocai Shan, Liyong Ye, Peng Feng, Xuejun Yan, Chenfan Li, and Yifan Zhao. An adaptive pig face recognition approach using convolutional neural networks. *Computers and Electronics in Agriculture*, 173:105386, 2020. 2

[29] Su Myat Noe, Thi Thi Zin, Pyke Tin, and Ikuo Kobayashi. Automatic detection and tracking of mounting behavior in cattle using a deep learning-based instance segmentation model. *International journal of innovative computing, information & control: IJICIC*, 18:211–220, 02 2022. 1

[30] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. C. Burie, C. L. Liu, and J. M. Ogier. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition. In *ICDAR*, 2019. 3

[31] Nvidia. Nvidia Deepstream SDK. 4, 5

[32] NVIDIA-AI-IOT. DeepStream Python Apps. 5

[33] Fumio Okura, Saya Ikuma, Yasushi Makihara, Daigo Muramatsu, Ken Nakada, and Yasushi Yagi. Rgb-d video-based individual identification of dairy cows using gait and texture analyses. *Computers and Electronics in Agriculture*, 165:104944, 2019. 2

[34] Yongliang Qiao, Daobilige Su, He Kong, Salah Sukkarieh, Sabrina Lomax, and Cameron Clark. Individual cattle identification using a deep learning based framework**the authors acknowledge the support of the meat livestock australia donor company through the project: Objective, robust, real-time animal welfare measures for the australian red meat industry. *IFAC-PapersOnLine*, 52(30):318–323, 2019. 6th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL 2019. 2

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 3

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. In *arXiv 1804.02767*, 2018. 3

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[38] L. Riaboff, A. Relun, C.-E. Petiot, M. Feuilloy, S. Couvreur, and A. Madouasse. Identification of discriminating behavioural and movement variables in lameness scores of dairy cows at pasture from accelerometer and gps sensors using a partial least squares discriminant analysis. *Preventive Veterinary Medicine*, 193:105383, 2021. 1

[39] Luis Ruiz-Garcia and Loredana Lunadei. The role of rfid in agriculture: Applications, limitations and challenges. *Computers and Electronics in Agriculture*, 79(1):42–50, 2011. 1

[40] Parvinder Singh and Amandeep Kaur. Chapter 2 - a systematic review of artificial intelligence in agriculture. In Ramesh Chandra Poonia, Vijander Singh, and Soumya Ranjan Nayak, editors, *Deep Learning for Sustainable Agriculture*, Cognitive Data Science in Sustainable Computing, pages 57–80. Academic Press, 2022. 1

[41] Joris Vandermeulen, Claudia Bahr, Dayle Johnston, Bernadette Earley, Emanuela Tullo, Ilaria Fontana, Marcella Guarino, Vasileios Exadaktylos, and Daniel Berckmans. Early recognition of bovine respiratory disease in calves using automated continuous monitoring of cough sounds. *Computers and Electronics in Agriculture*, 129:15–26, 2016. 1

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[43] Maxime Vidal, Nathan Wolf, Beth Rosenberg, Bradley P Harris, and Alexander Mathis. Perspectives on Individual Animal Identification from Biology and Computer Vision. *Integrative and Comparative Biology*, 61(3):900–916, 05 2021. 2

[44] N. Volkmann, B. Kulig, S. Hoppe, J. Stracke, O. Hensel, and N. Kemper. On-farm detection of claw lesions in dairy cows based on acoustic analyses and machine learning. *Journal of Dairy Science*, 104(5):5921–5931, 2021. 1

[45] Zhuang Wan, Fang Tian, and Cheng Zhang. Sheep face recognition model based on deep learning and bilinear feature fusion. *Animals*, 13(12), 2023. 2

[46] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *ECCV*, 2018. 3

[47] Zhang Xudong, Kang Xi, Feng Ningning, and Liu Gang. Automatic recognition of dairy cow mastitis from thermal images by a deep learning detector. *Computers and Electronics in Agriculture*, 178:105754, 2020. 1

[48] Liyao Yao, Zexi Hu, Caixing Liu, Hanxing Liu, Yingjie Kuang, and Yuefang Gao. Cow face detection and recognition based on automatic feature extraction algorithm. In *Proceedings of the ACM Turing Celebration Conference - China*, ACM TURC '19, New York, NY, USA, 2019. Association for Computing Machinery. 2

[49] Qixiang Ye and David Doermann. Text detection and recognition in imagery: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(7), 2015. 3

[50] Harsha Yelchuri and Rashmi R. A review of tinyml, 2022. 2

[51] Sun Yukun, Huo Pengju, Wang Yujie, Cui Ziqi, Li Yang, Dai Baisheng, Li Runze, and Zhang Yonggen. Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score. *Journal of Dairy Science*, 102(11):10140–10151, 2019. 2

[52] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. *ICCV*, 2017. 3

[53] Thi Thi Zin, Shuhei Misawa, Moe Zet Pwint, Shin Thant, Pann Thinzar Seint, Kosuke Sumi, and Kyohiro Yoshida. Cow identification system using ear tag recognition. In *2020 IEEE 2nd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 65–66, 2020. 2

[54] Thi Thi Zin, Moe Zet Pwint, Pann Thinzar Seint, Shin Thant, Shuhei Misawa, Kosuke Sumi, and Kyohiro Yoshida. Automatic cow location tracking system using ear tag visual analysis. *Sensors*, 20(12), 2020. 2