Supplementary Material: MmCows: A Multimodal Dataset for Dairy Cattle Monitoring

Hien Vu
Purdue University
hienvu@purdue.edu

Omkar Prabhune
Purdue University
oprabhun@purdue.edu

Unmesh Raskar
University of Wisconsin–Madison
uraskar@wisc.edu

Dimuth Panditharatne
University of Wisconsin–Madison
panditharatn@wisc.edu

Hanwook Chung
Iowa State University
hwchung@iastate.edu

Christopher Y. Choi
University of Wisconsin–Madison
cchoi22@wisc.edu

Younghyun Kim
Purdue University
younghyun@purdue.edu

Overview

This document provides additional details that complement the main paper. We discuss the steps used to synchronize and calibrate the visual data in Section A. Section B elaborates on the details of UWB localization, heading direction estimation, and obtaining the reference for lying behavior. In Section C, we explain the rationale for defining the ground truth, as well as providing a detailed derivation of visual localization. In Section D, we provide further details on benchmarks and implementation, along with analyses of the experimental results. Finally, we discuss the ethical considerations, utilization, generalizability, and limitations in Section E, as well as an additional visualization tool in Section F.

We keep the order of figures, tables, and equations in numerical, and refer to them independently from the main paper unless explicitly stated otherwise. The sections in this document are kept in alphabetical order.

The paper checklist is attached as the final part of the main paper. The dataset and the code for benchmarks are available at `https://github.com/neis-lab/mmcows`

A Visual Data Processing

We discuss additional details of processing the visual data and calibrating four camera views.

Visual data synchronization and timestamp alignment. We synchronize image frames from independent cameras using Internet time. During the deployment, we periodically placed an Internet-time-synchronized clock in front of the camera views, where the locations of the frames that captured the clock are pinpointed during post-processing for frame alignment and timestamp synchronization. The cameras recorded the Internet time every two to three days, totaling 16 clock records for all cameras. To ensure the high quality of the images, we turned off the cameras for a few minutes each time to clean the lens to prevent dust from building up. The synchronization satisfied three requirements: the resulting timestamps of the frames match exactly with the clock information in the images, the total number of available and missing frames in a single day is equal to the total number of seconds in one day, and the timestamps of light-turning-on and -off events in the pen at night remain consistent throughout the deployment. After the synchronization, the recorded videos were extracted as images and cropped from 5.1k to 4.5k.

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

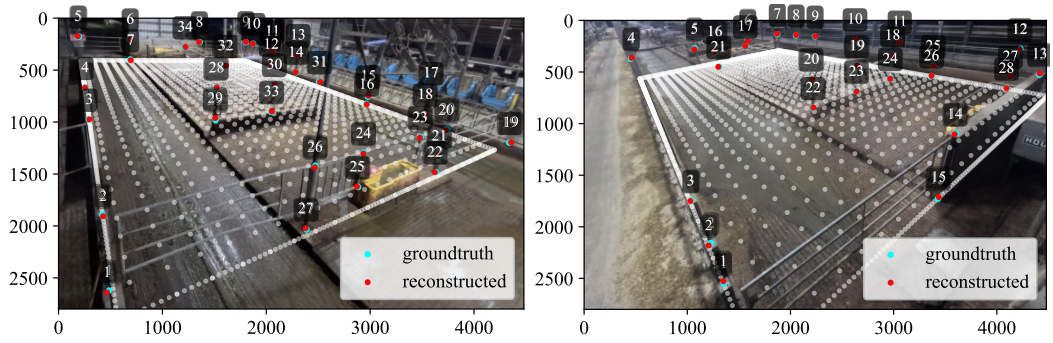


Figure 1: Calibration for Cameras #1 (left) and #3 (right) with the ground truth and reconstructed points that are projected from the 3D world to the camera views along with the ground grid (the axes represent pixel locations).

Since the ultra-wideband (UWB) distance measurements are synchronized and the timestamps are known, we also maintain a separate set of images that are captured at the same time as the UWB measurements throughout the deployment, with a sampling interval of 15 seconds, called UWB-synced frames.

Camera calibration. As location data is an important part of this dataset, we also provide projection matrices for transforming 3D points in the world coordinates to 2D points on the camera views and vice versa. The projection matrices are derived from direct linear transformation using 3D coordinates of 33 points measured around the pen and their corresponding 2D locations of pixels in each camera view. The 3D measured points and the corresponding reconstructed points, along with the ground grid, are illustrated in Figure 1.

B Obtaining the Secondary Data

The measured data, such as UWB distances, magnetic field, and ankle acceleration, require additional processing steps to extract meaningful data for inferring cows' activities. In this section, we describe the processing methods applied.

B.1 3D localization using UWB

UWB-based locations are retrieved from the distances measured between each tag and multiple stationary anchors. To perform the UWB distance measurements, the tag initiates a two-way ranging session with each of the eight anchors consecutively. In each ranging session between a tag and an anchor, the tag performs five measurements for oversampling. In addition to the distance from the tag to the anchor, other parameters are also collected, such as the number of successful measurements, the average line-of-sight probability, and the average received signal strength.

From eight stationary UWB anchors, we define two groups of anchors for localization based on their IDs: [1,2,4,6,7] and [2,3,5,7,8]. When calculating the location of a cow, only the group with a shorter total distance to the tag is selected for localization.

In order to compute the optimal location of a cow using UWB distances, the 3D location is iteratively computed using an optimization-based localization approach. To find the location of a mobile tag in a positioning system with N anchors, let (x_t, y_t, z_t) represent the location of the tag, and (x_i, y_i, z_i) represent the location of the i-th anchor among N stationary anchors. The position of the tag is determined by minimizing the loss function:

$$f(x_t, y_t, z_t) = \sum_{i=1}^{N} \left(\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2} - d_i \right)^2 \tag{1}$$

where d_i is the distance between the tag and the i-th anchor. To minimize the loss function, the gradient is determined based on the following partial derivatives of the loss function:

$$\frac{\partial f}{\partial x_t} = f_x(x_t, y_t, z_t) = \sum_{i=1}^{N} \frac{2(x_t - x_i)(\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2} - d_i)}{\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2}} \tag{2}$$

$$\frac{\partial f}{\partial y_t} = f_y(x_t, y_t, z_t) = \sum_{i=1}^{N} \frac{2(y_t - y_i)(\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2} - d_i)}{\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2}} \tag{3}$$

$$\frac{\partial f}{\partial z_t} = f_z(x_t, y_t, z_t) = \sum_{i=1}^{N} \frac{2(z_t - z_i)(\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2} - d_i)}{\sqrt{(x_t - x_i)^2 + (y_t - y_i)^2 + (z_t - z_i)^2}} \tag{4}$$

We then use AdaGrad [1] with the recursive functions as follows:

$$x_n = x_{n-1} - \frac{\eta f_x(x_n, y_n, z_n)}{\sqrt{\sum_{i=1}^{n} f_x(x_i, y_i, z_i)^2}} \tag{5}$$

$$y_n = y_{n-1} - \frac{\eta f_y(x_n, y_n, z_n)}{\sqrt{\sum_{i=1}^{n} f_y(x_i, y_i, z_i)^2}} \tag{6}$$

$$z_n = z_{n-1} - \frac{\eta f_z(x_n, y_n, z_n)}{\sqrt{\sum_{i=1}^{n} f_z(x_i, y_i, z_i)^2}} \tag{7}$$

where η is the step size. After a certain number of iterations, the optimal location of the tag is obtained.

The accuracy of UWB 3D localization is evaluated using two reference tags, based on the Circular Error Probability at the 95% level (CEP-R95), which is defined for each tag as the radius of a circle centered on the mean position of all 3D locations in which 95% of the locations fall within [2]. The CEP-R95 of the two reference tags (ID: #13 and #14) throughout the experiment is 8.31 cm and 8.44 cm.

B.2 3D direction of the cow's head

For the first time, we provide the head direction that complements the neck location to offer more insights into cattle behaviors. As the cows' movements are very slow and they spend most of their time staying still, we assume the acceleration vector recorded by the inertial and magnetic measurement unit (IMMU) represents the direction of the Earth's gravity, which is vertical in the 3D coordinate. Additionally, the direction of the Earth's magnetic field is also consistent throughout the deployment. As the gravity and magnetic directions are non-parallel and consistent, they are used as reference vectors for the cow's head direction. Data from the IMMU are used to compute the head direction in roll, pitch, and yaw. We use the tilt-compensated eCompass [3] to extract the cow's head direction. The heading direction was calculated in the North-East-Down (NED) coordinate system and subsequently converted to the East-North-Up (ENU) coordinate system to remain consistent with UWB and visual locations.

Due to the geographic location of the deployment site being at a high latitude of 43°N 89°W, the magnetic vector points downward at a large inclination angle of 69° [4], leaving the relative angle between the two reference vectors at about 21°. As a result, any distortion of the magnetic field occurs when the sensor is too close to a big metal structure like the feed lock, or any sudden movement of the cow could reduce the relative angle between the two vectors, resulting in a wrong estimation of the head direction. Therefore, the head direction data should not be used when the relative angle is smaller than a certain value, e.g. 10°. When the relative angles smaller than 10° are filtered out, the availability of the head direction data from ten neck tags varies from 82% to 96%.

B.3 The reference for lying behavior

In precision livestock farming (PLF), using the accelerometer to measure the lying duration of swine and cattle is a common approach [5, 6]. Ten out of 16 cows are equipped with ankle sensors that measure three-axis acceleration, which indicates the direction of the Earth's gravity. The ankle sensors were mounted such that the y-axis points downward and is parallel to the leg's orientation. When the cow stands upright, this axis will have maximum readings at around 1 g. When the cow is lying, either on the right or left side, the readings on the axis will be at a minimum. The maximum and minimum

Figure 2: Examples of seven behaviors of the cows: 1-walking, 2-standing, 3-feeding head up, 4-feeding head down, 5-licking, 6-drinking, and 7-lying, with 0 indicates the behavior is unknown

readings form two clusters, where K-Means clustering is used to determine a middle threshold. The cow is detected as lying when the reading is smaller than the threshold and non-lying otherwise. When compared to manually annotated lying behavior ground truth in Section C, the average accuracy of ankle-based lying detection from ten cows in one day is 99.75%. The effectiveness of this approach allows the ankle acceleration to be used as a reference for the lying behavior during 14 days of the deployment.

C Further Details of the Ground Truth

We discuss further details on how the cow ID and behavior ground truth were created.

C.1 Ground truth for cow identification

We selected UWB-synced frames from four cameras on July 25th from 2:57:18 AM to 11:57:17 PM, when the lighting was available for cow ID annotation and behavior labeling, totaling 20,000 frames. The annotation rules focus on cow identification in such a way that we only annotate a non-lying cow if there is a visible portion of the cow's body with a recognizable pattern that is significant enough for the cow ID to be deemed identifiable. When the cows are lying in the stalls, their bodies are often heavily occluded by one another. This makes it more challenging to identify the cow as a model cannot rely on the body pattern to identify the cow. In this case, tracking the cow from a standing position as it transitions to lying might help identify the cow better. As a result, we annotate the cows as long as their body shape is visible so that the cow can be detected accurately. Out of a total of 213,000 bounding boxes of lying and non-lying cows, the number of bounding boxes for each cow varies from 10,000 to 15,000. On average, there are 10.6 annotated cows per frame.

As the cows are observed at different viewing angles and directions in their natural positions, identifying cows is more challenging compared to the case when using top-view images of cows. In many instances, one standing cow could be occluded by one or a few other cows, and the size of the cow appears differently in each camera view. The annotators were trained to follow our strict annotation rules to ensure consistency. The details and visual examples of the annotation rules will be made available on the dataset website. We used the VGG Image Annotator (VIA) to annotate the cow IDs [7, 8].

C.2 Ground truth of individual cows' behaviors

For labeling cow behaviors, we define seven behaviors: walking, standing, feeding head up, feeding head down, licking, drinking, and lying. The definitions, visual examples, and statistics of the behaviors are provided in Table 1, Figure 2, and Table 2, respectively. The behavior is labeled as zero when the cow is not visible in any camera view, which only happens when she leaves for milking twice per day or when the light is off from 11:57:19 PM to 2:57:17 AM.

Table 1: Behavior definitions of cows observed in isometric views

#	Behavior	Definition
0	Unknown	When the cow is not visible in any camera view
1	Walking	Moving from one location to another between consecutive frames
2	Standing	The legs are straight up, and the head is not at the feeding area
3	Feeding head up	The head is at the feeding area and the mouth is above the food
4	Feeding head down	The head is at the feeding area and the mouth touches the food
5	Licking	Licking the mineral (salt) block
6	Drinking	Drinking at a water trough, when the mouth touches the water
7	Lying	The cow is lying in the stall

Table 2: Statistics of cow behaviors in one-day ground truth data at two different sample rates

Sampling rate	Behaviors						
	Walking	Standing	Feeding↑	Feeding↓	Drinking	Licking	Lying
1 s (full ground truth)	1.48%	25.21%	7.00%	11.41%	0.70%	1.23%	52.98%
15 s (UWB synced)	1.43%	25.16%	6.93%	11.49%	0.70%	1.21%	53.07%

During feeding, the cows also stand, but they are labeled as feeding for the purpose of behavior separation. We separate feeding behavior into 'feeding head up' and 'feeding head down' because sometimes cows do not feed even when their heads are above the food. It has been shown that cows only feed from 68% to 88% of the time while staying in the feeding area [9]. When the head is above the food, the cow could be chewing, and this behavior separation can be used to obtain a more fine-grained feeding duration such that the cow can be assumed to feed while the head is above the food for a certain number of seconds following the last moment when the cow took in the food with its head down.

The labeling process is thoroughly conducted to ensure accuracy and consistency across behavior labels generated by multiple annotators before, during, and after the labeling process. Before labeling, the annotators underwent an extensive training process where they distinguished between true positive and false positive cases of each behavior from various viewing angles. They were required to pass a screening test in which they had to track and assign correct behaviors of a cow that moves between other cows around the pen for a long duration. During labeling, the annotated cow IDs are added to the footage at which they show up every 15 frames to help the annotator accurately track the cow of interest. After labeling, another annotator is assigned to cross-check the labels to ensure that a cow only switches between behaviors that can only be performed at the same location–between drinking and standing, between feeding head up and feeding head down, and between licking and standing.

C.3 Visual localization and location ground truth

As locations are important in inferring cows' behaviors, our goal is to provide a reliable location ground truth. Since the bounding boxes and IDs have been manually and thoroughly verified during the annotation, they can be used to extract the most accurate locations of the cows. We define the location ground truth as the central location of the cow's body, which is typically observed at the center of the annotated bounding box in each camera view. We propose a new optimization-based approach to calculate the 3D body location using annotated bounding boxes of the same cow in multiple camera views. The location is derived by projecting the bounding box centers to the world coordinate system as 3D lines that inherently converge. We then apply AdaGrad to find the optimal location, which is nearest to the lines, resulting in the 3D body location.

To find this 3D point, from the projection matrix of a camera $P = [M|p_4]$, the camera location in the 3D world is $C_0 = -M^{-1}p_4$ [10]. The projection line of the bounding box center is represented as

$$C = C_0 + tv \tag{8}$$

where C is any point on the line, v is the direction vector of the line, and t is the scaling factor. Given point Q outside of the line, we need to find the distance from Q to the line. Assuming the nearest

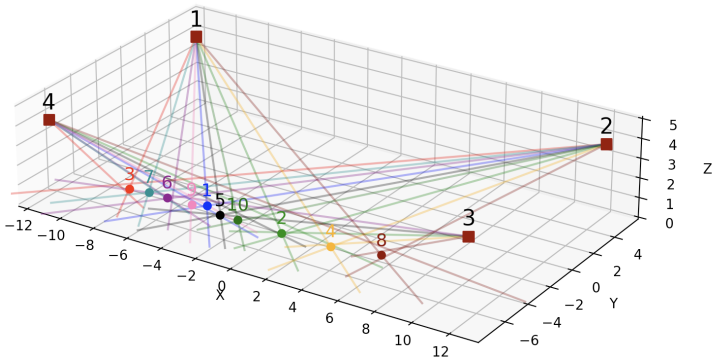


Figure 3: 3D visual localization of multiple cows using multi-view images. The lines with the same color represent the centers of bounding boxes of the same cows in different views.

point on the line to Q is K, this point can be found by projecting $\overrightarrow{C_0Q}$ on the line:

$$t_0 = \frac{\overrightarrow{C_0Q} \cdot \vec{v}}{||\vec{v}||^2} \tag{9}$$

$$K = C_0 + t_0 v, \tag{10}$$

which can be used to calculate the distance from point Q to the line, $d(Q,K)$. With N lines from N camera views where the bounding box centers are visible, the objective function that needs to be minimized is:

$$f(x,y,z) = \sum_{i=1}^{N} d(Q,K_i). \tag{11}$$

To minimize the total distance from point Q to the lines, point Q needs to move in the direction of $\overrightarrow{QK_i}$ $(1 \leq i \leq N)$. In this case, the total gradient is:

$$\nabla f(x,y,z) = \sum_{i=1}^{N} \frac{\overrightarrow{QK_i}}{||\overrightarrow{QK_i}||}. \tag{12}$$

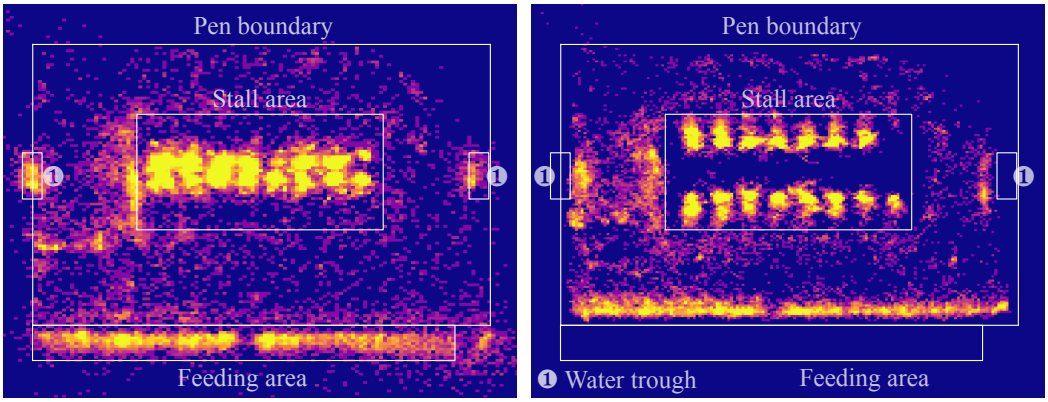
By applying AdaGrad [1], after each iteration, Q will move closer to the optimal location of the bounding box centers in the 3D world coordinate:

$$Q_n = Q_{n-1} - \frac{\eta \nabla f(x_n, y_n, z_n)}{\sqrt{\sum_{i=1}^{n} \nabla f(x_i, y_i, z_i)^2}} \tag{13}$$

where η is the step size. The final location is used as the ground truth of the cow's body location. Figure 3 illustrates the projected lines from four cameras to the 3D world and the optimal locations of multiple cows.

This approach requires the cow to be visible in at least two camera views, which is true for all cows 87.5% of the time on the chosen day July 25$^{\text{th}}$. When the cow is only visible in one camera view (12.5% of the time, which mostly happens when the cows are lying), the location is calculated by projecting the center point to the 3D world coordinate at an assumed height. The assumed height is set differently for lying and non-lying cows. We find the most representative values of the z-direction of the visual locations for lying and non-lying cows 55 cm and 80 cm, respectively, calculated using at least two camera views. The values are used as the assumed height depending on whether the cow is lying or non-lying. Only 2.5% of the time is the bounding box of a cow not available in any camera view. The cows are available 22.5%, 27.4%, and 35.1% of the time in two, three, and four camera views, respectively.

Figure 4 shows the heat maps of UWB neck locations and visual body locations of the cows on July 25$^{\text{th}}$. The body locations of the cows are perfectly located inside the pen area except for a few data points at the top-left corner where the cows move out of the pen for milking, while their locations are



UWB neck locations Visual body locations

Figure 4: Heat maps of 54,000 UWB neck locations of 10 cows (left) in 23 hours and 74,000 visual body locations of 16 cows (right) in 20 hours on July 25th (no data during two milking sessions that totaled one hour for both types, and during three hours of light off for visual locations)

clearly separated in the stall area where they stand or rest, demonstrating the high accuracy of visual localization. The UWB neck locations are reliable enough as the cows can stick their neck out of the pen during feeding, except for data points outside of the feeding area, which we discuss further in Section E.

For multi-view multi-cow visual localization using vision models, the 3D visual body location can be converted to an indexed-grid location by removing the z-axis and assigning it to the corresponding 2D indexed ground grid. The newly indexed location can be used as the ground truth for training and benchmarking the localization accuracy of multi-view visual localization models.

D Details of Benchmark and Implementation

D.1 Comparison of modalities for behavior classification

In this section, we describe in detail the implementation of the models and result analyses. We utilized PyTorch, TensorFlow, and Scikit-learn to implement the models [11, 12, 13]. Intel Xeon Gold 6138 CPU @ 80x 2GHz and NVIDIA Tesla V100 SXM2 4x are used. The training code and pre-trained model will be made open source. Both the primary data and secondary data will be released to the public. The dataset will be made available under the following Creative Commons license: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) [14]. The code will be made available under the MIT open-source license [15].

Evaluation metrics. We discuss further how the F1 score is calculated in RGBs and RGBm, as they involve cow identification. Instead of evaluating the performance of a single behavior classification model, we evaluate the F1 scores of the whole pipeline, where both the cow ID and its behavior need to be predicted accurately.

In the case of RGBs, considering a single frame, the model is expected to correctly detect, classify, and identify the visible cows in the frame. The results are then compared to the annotated IDs and behavior labels, i.e., the available bounding boxes and IDs in the ground truth of the frame, as well as the corresponding ground truth behavior labels. For every cow that is available in the ground truth, if it is also available among the predicted cows, the corresponding predicted behavior label will be compared against the ground-truth behavior label for computing the F1 scores. If the cow is not found among the predicted cows, the predicted behavior is considered unknown (denoted as 0). We discard all predicted cows that are not present among the cows in the ground truth.

For RGBm, the list of available cows in the ground truth is combined from all four annotated labels from four camera views that were taken at the same time. We then perform the same process as for the RGBs to calculate the F1 scores. Computing the F1 score in this way ensures that if the

Table 3: Ten groups of data using temporal split (TS) in the first fold of cross-validation

Group	Train			Val	Test
Artificial light	{1} 2:57–4:30	{2} 4:30–6:00	{3} 18:00–20:00	{4} 20:00–22:00	{5} 22:00–23:57
Natural light	{6} 6:00–8:24	{7} 8:24–10:48	{8} 10:48–13:12	{9} 13:12–15:36	{10} 15:36–18:00

Table 4: Performance of the four vision models in RGBs

Model	$\text{mAP}_{[0.5:0.95]}$ ↑	Average F1 ↑
Cow detection	$.729_{\pm.014}$	-
Cow detection	$.729_{\pm.014}$	-
Behavior classification	-	$.678_{\pm.025}$
Lying cow identification	-	$.540_{\pm.060}$
Non-lying cow identification	-	$.942_{\pm.011}$

model cannot detect the cow, the predicted behavior label is assigned zero even though there is no information about the cow being aggregated through the pipeline.

Split settings. In object-wise split (OS), ten cows are grouped into pairs based on cow IDs {1,2}, {3,4}, {5,6}, {7,8}, and {9,10}. In the first fold of cross-validation, the cow IDs for training, validation, and testing are {1,2,3,4,5,6:7,8:9,10}. We shift the order by one pair in each fold such that in the second fold, the order is {3,4,5,6,7,8:9,10:1,2}, and so on. In some models where the validation is not needed, the validation set is concatenated to the test set.

In the temporal split (TS), because of the differences in the appearance of cows in the RGB data under different lighting conditions, the data from each modality is divided into two groups based on two lighting conditions: artificial light and natural light. The data from each modality is separated into ten chunks ({1} to {10}) as shown in Table 3. In the first fold of cross-validation, the training, validation, and testing sets based on chunk IDs are {1,2,3,6,7,8:4,9:5,10}. We then shift by two chunks in each subsequent fold such as {2,3,4,7,8,9:5,10:1,6} and so on.

To ensure that the algorithms do not learn cow-specific behaviors, OS is used that excludes two cows from the first ten cows in each validation fold for behavior classification. In other words, this is similar to leave-one-cow-out cross-validation. We use OS for non-vision data, but not for vision data, because our vision pipeline performs identification and behavior classification at the same time, making it difficult to apply OS validation. The cow identification models are trained using only TS as the data from all cows are required. A different vision pipeline with separate identification and behavior classification could be designed, where behavior classification is completely ID-agnostic, but we leave this potential approach out of the scope of this work.

UWB. We use a Random Forest (RF) model with balanced weights, where the classes are weighted inversely proportional to their popularity in the data. UWB data at 15 seconds intervals are used, where each location point is associated with a specific behavior. In the UWB's results, the average F1 score is about .712 with low error rates. The drinking score is lower while the error rate is higher, which can be explained by the fact that the cows often stand with their neck above the water trough but do not always drink, creating confusion between standing and drinking.

IMMU. We truncated the acceleration data into 10-second windows with an overlap of 50%, making each data sample 5 seconds apart. Discrete wavelet transform is used to extract the approximation and detail coefficients of the acceleration [16]. The wavelet coefficients are concatenated with the relative angle between the acceleration and magnetic vector, which is derived from the head direction data. The data includes four features, which are normalized separately. The behavior label is selected from the timestamp at the middle of each window. We employ a fully connected network that consists of four layers with 300, 256, 127, and 7 neurons. All activation functions are ReLu except for the final layer. A dropout layer and batch normalization are added after each fully connected (FC) layer (excluding the final layer) to reduce the dependency on specific neurons.

We conducted an independent test of IMMU using a random split of data from a single cow that showed an average F1 score of .605 which is comparable to UWB. However, IMMU performs worst

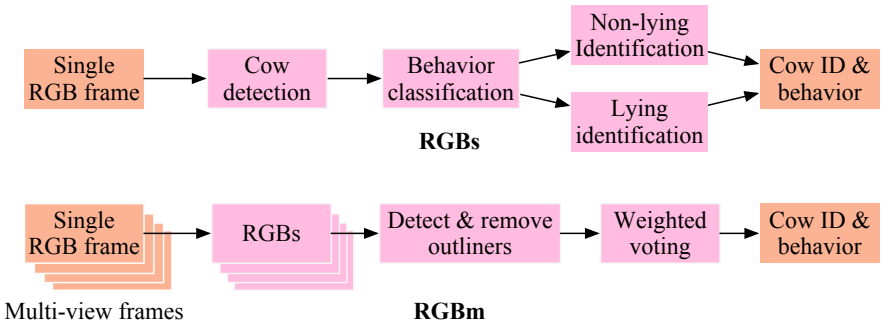


Figure 5: Processing pipelines of RGBs (top) and RGBm (bottom) for visual cow identification and behavior classification

in both OS and TS despite having more features. This low performance shows the poor transferability of IMMU among cows.

RGBs. As shown in Figure 5, the computer vision-based behavior monitoring pipeline consists of 3 stages: cow detection, behavior classification, and cow identification.

In the first stage, we employ YOLOv8 [17] as an object detector to detect cows in a video frame. YOLOv8 predicts the bounding box coordinates of cows that are present in the image. The model was trained for 20 epochs using the default training parameters provided in [17]. To ensure robustness and generalizability, we used temporal split (TS) as described previously and trained five YOLOv8 models based on five different timings of train, validation, and test splits, achieving a mean Average Precision (mAP) of $.729_{\pm.014}$, as shown in Table 4.

Once the bounding boxes of cows are predicted, the detected cows are cropped from the image and fed into the behavior classifier. The behavior classifier is a convolutional neural network (CNN) tasked with classifying the predicted cows into one of seven behaviors, namely walking, standing, feeding head up, feeding head down, drinking, licking, and lying. We use EfficientNet-B0 [18] for this task due to its balance between performance and computational efficiency. The cropped detections that form the input to the classification model are resized with padding to 224×224 and normalized using the mean and standard deviation computed using the training data split. The model is trained using a categorical cross-entropy loss function and Adam optimizer with a learning rate of .001 for 15 epochs. The same five-fold temporal split (TS) configurations were used to train and test the model, and an F1 score of $.678_{\pm.025}$ was obtained for the behavior classification task when tested independently on the test split of behavior classification data. Table 5 shows the class-wise performance of the behavior classification model.

The detected cows cropped from the original images are classified into one of 16 classes, where each class represents an individual cow identity. For this stage, we train two separate image classification models: one to predict the identity of non-lying cows and another for lying cows. The model architecture used for cow identification, i.e., EfficientNet-B0 [18], is the same as that for behavior classification. Using two different models instead of a single model accounts for the significant differences in appearance between non-lying and lying cows, which could impact the model's ability to accurately differentiate between individual identities. The model is trained using a categorical cross-entropy loss function and Adam optimizer with a learning rate of .001 for 15 epochs. The model was trained and tested using the same five-fold temporal split (TS) configurations as the previous two stages. F1 scores of $.540_{\pm.060}$ for lying cows and $.942_{\pm.011}$ for non-lying cow identification were obtained when the models were tested independently using test data for cow identification. Table 6 shows the ID-wise performance of the cow identification models. During inference, the behavior classification model's prediction determines which cow identification classifier to use. If the predicted behavior is 'lying', the lying cow classifier is used. For all other behaviors, the non-lying cow classifier is used.

UWB+HD. We combine the `uwb` and head direction (`hd`) to create the training data, where `hd` is downsampled to a sampling rate of 15 seconds. We also use Random Forest (RF) with balanced weights that result in a slight improvement in all behaviors compared to the performance of UWB.

Table 5: Performance of the behavior classification model

	F1 score ↑						
	Walking	Standing	Feeding↑	Feeding↓	Drinking	Licking	Lying
All cows	.179±.062	.901±.030	.704±.040	.819±.024	.591±.161	.569±.066	.986±.007

Table 6: Performance (F1 score) of the two visual cow identification models

Cow ID	Non-lying cow identifier	Lying cow identifier
1	.954±.010	.584±.142
2	.956±.021	.566±.163
3	.958±.008	.656±.231
4	.966±.031	.568±.193
5	.968±.019	.533±.276
6	.901±.026	.240±.192
7	.935±.012	.458±.269
8	.967±.013	.570±.301
9	.961±.024	.326±.218
10	.948±.033	.476±.119
11	.833±.051	.702±.111
12	.975±.008	.580±.190
13	.977±.009	.802±.090
14	.899±.031	.438±.092
15	.967±.017	.448±.150
16	.906±.026	.622±.125

UWB+HD+Akl. We use the same data format and model as in UWB+HD but incorporate the additional `ankle` data which was up-sampled from 1-minute to 15-second intervals. The results showed improvements in all behaviors with lower error rates.

RGBm. We use multiple views of the cow at the same time instance to predict the behavior. Each frame is processed by RGBs to produce bounding boxes with cow IDs. For each predicted bounding box of the same cow, the center is projected to the world coordinate as a 3D line. We use visual localization to find and exclude the projection line that does not converge with other lines. We then implement weighted voting to determine the final cow's behavior where the weight is proportional to the width of the bounding box.

RGBm demonstrated a noticeable improvement from RGBs in all behavior classes as knowledge from multiple views is combined, while visual localization ensures that the same cow is being asserted among the views. Even though the performance of RGBm was not as high as UWB, it still holds great potential in advancing precision livestock farming (PLF) with the benefits of being low-cost, animal-friendly, and scalable.

D.2 Behavior analysis

We used UWB+HD+Akl to extract the behaviors of ten cows throughout the deployment. An example of a cow's behavior changes is shown in Figure 6, where the daily total duration of standing fluctuated in the same trend as the THI.

Method. The behaviors are pre-processed before being analyzed. For dairy cattle, the changes in a cow's position between behaviors for a short duration often create time gaps that segment the behavior into multiple sub-bouts, which can cause errors in calculating the number of bouts. To improve the accuracy, we used a custom moving-window filter that smooths out the behavior, such that the cow is considered as doing a certain behavior if the behavior is detected twice within a given window. A window size of 12 minutes is used for both standing and feeding behaviors, while a window of seven minutes is used for drinking. Figure 7 illustrates the smoothing process using this filter. The smoothed data is only used to calculate the number of bouts, whereas the duration of each bout and the total behavior duration are computed using the raw behavior data.

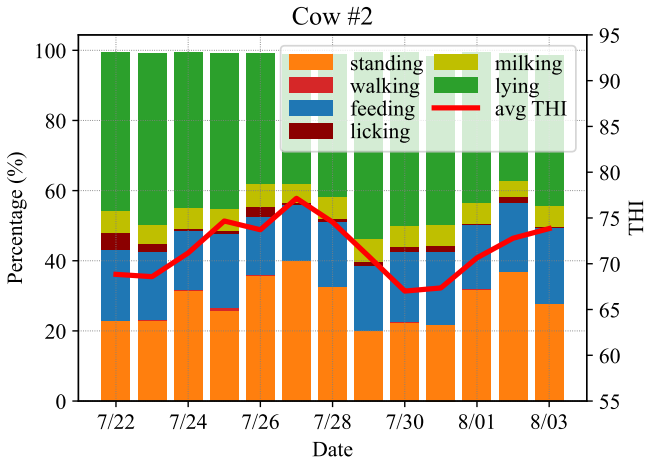


Figure 6: Behavior changes of cow #2 throughout the deployment where 100% corresponds to a duration of 24 hours

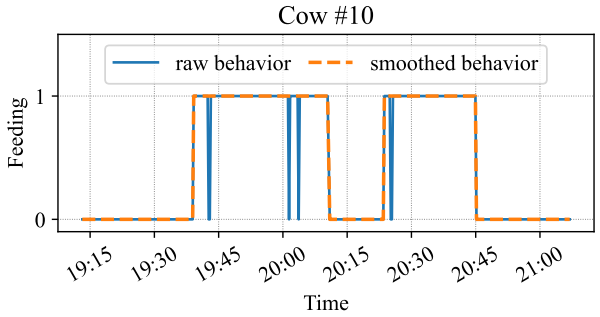


Figure 7: Feeding behavior of cow #10 before and after being smoothed out

Results. We confirm significant correlations between the cow behavior changes and THI as reported in Table 4 of the main paper. With r values of .726 and .678, the mean duration of standing bouts and total standing duration are strongly affected by the THI. Their correlations are also statistically significant, as the p-values are very small. The R^2 values, around .500, show that these parameters are predictable using THI, given the high natural variability between cows. These results confirm that cows stand longer when THI increases [19, 20, 21]. Similarly, the strong negative r values between lying and THI further verify that cows spent less time lying when THI was higher [22, 23].

Regarding feeding behaviors, we observe strongly correlated r values ($|r| > .500$) for feeding frequency and mean feeding duration. Given the low p-values, this suggests that cows often increase feeding visit frequency but reduce feeding duration during higher THI conditions, which is similar to findings in [24]. The strongest correlation that we found is between the drinking frequency and THI, with a very high r of .802 and a minimal p-value of .001. This result shows that dairy cows drink more when exposed to higher THI conditions, which aligns well with previous studies [25].

E Discussion

E.1 Ethical considerations

As the subjects of this dataset are dairy cattle, this dataset does not contain any personal information of any participant. To the best of our knowledge, this work does not negatively impact any person, animal, or entity during and after the deployment as well as after the release of MmCows to the public.

All sensor deployment procedures were done with the approval of the Institutional Animal Care and Use Committee (IACUC) of the University of Wisconsin–Madison (Protocol #A006606). All dairy cattle were handled in compliance with the European Directive 2010/63/EU [26] regarding the protection of animals used for scientific purposes.

E.2 Utilization and generalizability

So far, we have shown the benefits of MmCows in advancing precision livestock farming. Here, we further discuss the utilization of the dataset in specific research directions:

- **Feature Selection:** The multimodal nature of the dataset enables researchers to identify which combination of data modalities (e.g., visual, location, etc.) optimizes model performance for various outcomes. This analysis can improve model robustness across different applications where similar combinations of sensors are used, making the findings transferable to other contexts, such as adapting these models for different cattle breeds or farm setups.
- **Animal Identification:** The dataset presents challenges such as varying angles and distances in isometric views, occlusions between cows, and appearance differences across views. These challenges are ideal for developing algorithms for both closed-set and open-set identification of dairy cattle. The solutions derived from this dataset are not only applicable to Holstein cows but can also be adapted to closely related breeds, making the findings broadly transferable to different cattle populations.
- **Multi-View Camera Fusion:** The dataset supports research in fusing multi-view camera data, where detection and identification tasks require consolidation from multiple perspectives. This approach is particularly useful in complex environments where a single view may be obstructed or insufficient. The techniques developed using this dataset can be transferred to other livestock monitoring systems, ensuring improved detection and tracking accuracy in a variety of farm layouts and operational conditions.

Another important aspect of MmCows is the generalizability for different housing conditions and cattle breeds. Modalities that are less likely to be breed-dependent, such as `uwb`, `immu`, `pressure`, `ankle`, and `thi`, are expected to exhibit robust generalizability across different cattle breeds. On the other hand, visual data like `rgb` may have limited generalizability in visual identification for other breeds that do not have the Holstein-like pattern, but it can still be useful for visual localization and posture recognition. When comparing different housing conditions, such as indoor versus outdoor environments, the generalizability is expected to remain high for modalities that are not infrastructure-dependent, such as `immu`, `pressure`, `ankle`, and `thi`. However, modalities like `uwb` and `rgb` may be less applicable in outdoor settings, where obtaining location and visual data requires alternative sensors and devices, such as drones for capturing RGB images [27] or GPS for location tracking.

E.3 Data collection duration

The data collection was conducted over a period of two weeks instead of a longer duration which might be beneficial for various applications, particularly those that require tracking changes over extended periods. For the specific goal of behavior monitoring, our chosen two-week duration is sufficient to observe and accurately identify the seven targeted behaviors without requiring the more complex, expert-driven analysis needed for detecting long-term behavioral changes. Our focus was on establishing a robust foundation for behavior monitoring, where the effectiveness of short-term detection is critical. Moreover, the success of short-term monitoring lays the groundwork for future studies involving long-term data collection, which we plan to pursue as the next step in our research.

E.4 Sensor costs

The custom-designed neck-mounted tags in our dataset, which measure 3D location and head direction, are engineered to be cost-effective, especially when they are commercialized and mass-produced. With production costs potentially around 20–40 USD per unit, the device offers an affordable solution for large-scale deployment in dairy farms. Other commercial sensors, including ankle sensors and vaginal temperature sensors, are more costly, but they are only used for dataset generation and not for monitoring in the field. Most importantly, our multimodal dataset can assist system designers in evaluating the cost-effectiveness of different monitoring approaches, helping them choose the most

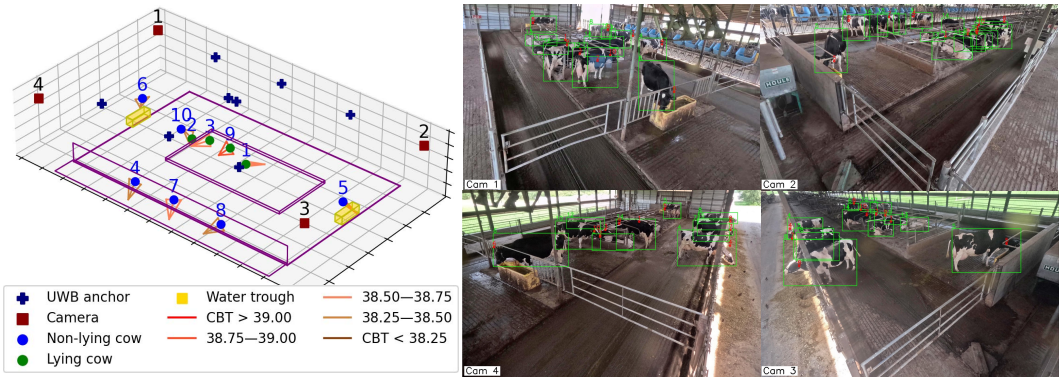


Figure 8: Data visualization tool for the MmCows with an interactive 3D map (left) and a combined camera view (right) with annotated bounding boxes in green and projected UWB locations in red

suitable sensors that balance performance and budget. By providing a detailed understanding of the trade-offs between sensor cost and monitoring capabilities, the dataset will enable more informed decision-making in herd management.

E.5 Limitations

Despite the large size and the diversity of modalities in MmCows, a few limitations remain. Due to limited resources, along with the complexity of handling large animals and system maintenance, the wearable sensors were only deployed on 10 cows. Nonetheless, the one-day visual and behavior ground truth data is available for all 16 cows. In addition, all UWB anchors were set to point inward toward the pen to ensure the best alignment between the UWB modules. This setup maximizes accuracy when the cows stay inside the pen, but does not always yield consistent results when the cow's neck is outside of the pen area, such as during feeding. A sensor network with many more additional UWB anchors could be used to address this limitation but at the cost of increased installation and maintenance complexity.

As one of the modalities in the dataset, pressure data could potentially be used to detect changes in the elevation of the cow's neck, providing valuable information on how often the cow stands up or lies down, which serves the same function as the ankle sensor in detecting standing/lying behavior. However, because the ankle sensor provides better accuracy, we did not include the pressure data in the evaluation section.

The dataset could also benefit future studies by including ground truth on the social behaviors of cattle. Social behavior–including the establishment of social hierarchy, competition for resources, and affiliative interactions–is known to profoundly influence feeding behavior, dry matter intake (DMI), and overall health. For instance, significant overcrowding can reduce feeding activity, alter resting behavior, and decrease rumination, while low-ranking individuals in a group may suffer from limited access to resources and increased aggression, leading to negative affective states and impaired health [28]. Although our dataset does not include ground truth for social behaviors, it provides a robust foundation for analyzing individual behaviors directly impacted by social dynamics. For example, changes in feeding or resting patterns within our dataset could indicate underlying social stress or competition within the herd [29]. Researchers could leverage this dataset to infer social interactions indirectly or to complement additional observational data specifically targeting social behavior.

The dataset can be further enhanced for early detection of lameness by incorporating locomotion scores assigned by an animal welfare specialist who reviews the multi-view footage. Lameness is often identified through subtle changes in gait and posture, which are critical to detecting early for timely intervention. By adding these scores, the dataset would not only provide a quantitative measure of lameness severity but also enable the development of predictive models for early detection. This enhancement would make the dataset a valuable resource for training machine learning (ML) models aimed at automated lameness detection, ultimately improving animal welfare and herd management practices by enabling continuous, real-time monitoring of cow health.

F Data Visualization Tool

For visualizing the multimodal data of MmCows from ten cows in all 14 days of the deployment, we provide an interactive 3D visualization tool that is intuitive and easy to use, as shown in Figure 8. The tool displays two time-synchronized windows separately: a 3D map and a combined camera view. The 3D map showcases various parameters such as the cows' location with the heading direction, the `cbt` which is illustrated through the color of the triangle that represents the cow's head, and the lying behavior as the color of the location point. The images in the combined view are displayed synchronously with the timestamp, which is chosen in the 3D map. Aside from showing the images, the combined camera view also provides options to illustrate the UWB locations on each image view, which is done through 3D projection. This helps the users to distinguish the cows from each other when the annotated bounding boxes are not available. Users also have multiple options to visualize more information on the images, such as adding the ground grid or the pen boundary, masking nearby cow pens, etc. This tool will be made available on the dataset website.

References

[1] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[2] Dean M. Anderson, Rick E. Estell, and Andres F. Cibils. Spatiotemporal cattle data—a plea for protocol standardization. *Scientific Research*, 2013.

[3] Talat Ozyagcilar. Implementing a tilt-compensated ecompass using accelerometer and magnetometer sensors. *Freescale Semiconductor, AN*, 4248, 2012.

[4] National Oceanic and Atmospheric Administration. Magnetic field calculators. Last accessed: October 30th, 2024. URL: `https://www.ngdc.noaa.gov/geomag/calculators/magcalc.shtml`.

[5] Jose M. Chapa, Kristina Maschat, Michael Iwersen, Johannes Baumgartner, and Marc Drillich. Accelerometer systems as tools for health and welfare assessment in cattle and pigs – A review. *Behavioural Processes*, 181:104262, 2020.

[6] Kim Margarette Corpuz Nogoy, Sun-il Chon, Ji-hwan Park, Saraswathi Sivamani, Dong-Hoon Lee, and Seong Ho Choi. High precision classification of resting and eating behaviors of cattle by using a collar-fitted triaxial accelerometer sensor. *Sensors*, 22(16):5961, 2022.

[7] Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the ACM International Conference on Multimedia (MM)*, 2019.

[8] Abhishek Dutta, Ankush Gupta, and Andrew Zisserman. VGG image annotator (VIA), 2016. Last accessed: October 30th, 2024. URL: `http://www.robots.ox.ac.uk/~vgg/software/via/`.

[9] Douglas D. Shane, Brad J. White, Robert L. Larson, David E. Amrine, and Jeremy L. Kramer. Probabilities of cattle participating in eating and drinking behavior when located at feeding and watering locations by a real time location system. *Computers and Electronics in Agriculture*, 127:460–466, 2016.

[10] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[11] PyTorch. Pytorch. Last accessed: October 30th, 2024. URL: `https://pytorch.org`.

[12] TensorFlow. Tensorflow. Last accessed: October 30th, 2024. URL: `https://www.tensorflow.org/`.

[13] Scikit-Learn. Scikit-learn. Last accessed: October 30th, 2024. URL: `https://scikit-learn.org/stable`.

[14] Creative Commons. Attribution-noncommercial 4.0 international. Last accessed: October 30th, 2024. URL: `https://creativecommons.org/licenses/by-nc/4.0/`.

[15] MIT. The MIT License, 2024. Last accessed: October 30th, 2024. URL: https://opensource.org/licenses/MIT.

[16] Ganapati Bhat, Ranadeep Deb, Vatika Vardhan Chaurasia, Holly Shill, and Umit Y Ogras. Online human activity recognition using low-power wearable devices. In *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–8. IEEE, 2018.

[17] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8, 2023. Last accessed: October 30th, 2024. URL: https://github.com/ultralytics/ultralytics.

[18] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[19] N. B. Cook, R. L. Mentink, T. B. Bennett, and K. Burgi. The effect of heat stress and lameness on time budgets of lactating dairy cows. *Journal of Dairy Science*, 90(4):1674–1682, 2007.

[20] J. D. Allen, L. W. Hall, R. J. Collier, and J. F. Smith. Effect of core body temperature, time of day, and climate conditions on behavioral patterns of lactating dairy cows experiencing mild to moderate heat stress. *Journal of Dairy Science*, 98(1):118–127, 2015.

[21] Grazyne Tresoldi, Karin E Schütz, and Cassandra B Tucker. Cooling cows with sprinklers: Effects of soaker flow rate and timing on behavioral and physiological responses to heat load and production. *Journal of Dairy Science*, 102(1):528–538, 2019.

[22] A. Gomez and N. B. Cook. Time budgets of lactating dairy cattle in commercial freestall herds. *Journal of Dairy Science*, 93(12):5772–5781, 2010.

[23] Lisette M. C. Leliveld, Elisabetta Riva, Gabriele Mattachini, Alberto Finzi, Daniela Lovarelli, and Giorgio Provolo. Dairy cow behavior is affected by period, time of day and housing. *Animals*, 12(4):512, 2022.

[24] J. Chang-Fung-Martel, M. T. Harrison, J. N. Brown, Richard Rawnsley, A. P. Smith, and Holger Meinke. Negative relationship between dry matter intake and the temperature-humidity index with increasing heat stress in cattle: A global meta-analysis. *International Journal of Biometeorology*, 65(12):2099–2109, 2021.

[25] Yu-Chi Tsai, Jih-Tay Hsu, Shih-Torng Ding, Dan Jeric Arcega Rustia, and Ta-Te Lin. Assessment of dairy cow heat stress by monitoring drinking behaviour using an embedded imaging system. *Biosystems Engineering*, 199:97–108, 2020.

[26] European Parliament and Council. Directive 2010/63/EU of the european parliament and of the council of 22 september 2010 on the protection of animals used for scientific purposes. *Official Journal of the European Union*, 276:33–79, 2010.

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

[28] R. J. Grant and J. L. Albright. Effect of animal grouping on feeding behavior and intake of dairy cattle. *Journal of Dairy Science*, 84:E156–E163, 2001.

[29] Margit B. Jensen. *The role of social behavior in cattle welfare*. Advances in Cattle Welfare, 2018.