Computers and Electronics in Agriculture 234 (2025) 110301

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Utilizing farm knowledge for indoor precision livestock farming: Time-domain adaptation of cattle face recognition

Shujie Han [a,b], Alvaro Fuentes [a,b], Jongbin Park [a,b], Sook Yoon [c,*], Jucheng Yang [d], Yongchae Jeong [e], Dong Sun Park [a,b,**]

[a] Department of Electronics Engineering, Jeonbuk National University, Jeonju 54896, South Korea
[b] Core Institute of Intelligent Robots, Jeonbuk National University, Jeonju 54896, South Korea
[c] Department of Computer Engineering, Mokpo National University, Muan 58554, South Korea
[d] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China
[e] Division of Electronics and Information Engineering, IT Convergence Research Center, Jeonbuk National University, Jeonju 54896, South Korea

ARTICLE INFO

Keywords:
Indoor cattle farm
Face detection
Face alignment
Face recognition
Farm knowledge

ABSTRACT

In real-field cattle farming environments, precise cattle recognition is imperative for effective animal husbandry practices such as monitoring individual behaviors and screening health to ensure animal welfare. Recently, data-driven deep learning models provide efficient and non-intrusive face recognition. However, their application in real-world scenarios presents significant challenges due to data domain drift over time, encompassing geometric variations in face pose, illumination fluctuations, and disruptions in the background environment. To tackle these challenges, this paper introduces a framework for cattle face recognition with innovative techniques based on farm knowledge that guides the model's training and inference process. First, we combine temporal and pose alignment to mitigate the impact of geometric pose variations. Second, we employ illumination augmentation to adapt to varying illumination conditions, bolstering model robustness. Third, we use semantic segmentation to isolate the facial components, enhancing recognition precision and maintaining focus on facial attributes. Empirical experiments validate our approach, demonstrating its effectiveness for real-world deployment, ensuring robust performance across changing environmental conditions. Our model maintains high accuracy, underscoring its reliability in managing the complexity of real-world scenarios. In summary, this paper presents a comprehensive strategy to address domain drift challenges in cattle face recognition within extended real-world settings, equipping the model to meet the demands of genuine cattle farming contexts effectively.

1. Introduction

Precision livestock farming (PLF) requires precise and accurate recognition of each individual animal within a barn to enable meticulous management (Li et al., 2020; Neethirajan, 2021). Commonly employed methods include Radio Frequency Identification (RFID) and computer vision algorithms utilizing surveillance cameras. Specifically, RFID devices such as ear tags or collars are widely used to record cattle identity, providing high accuracy but being constrained by environmental factors (Pretto et al., 2022; Qiao et al., 2021). These limitations primarily arise from signal interference and reliance on card reader equipment, which pose challenges for real-time system performance. In contrast, surveillance cameras offer a readily accessible solution for real-time monitoring, though their accuracy is not always consistent (Fuentes et al., 2023; Han et al., 2023; Qiao et al., 2023; Ter-Sarkisov et al., 2017).

Recent advancements in deep learning have significantly enhanced the accuracy of cattle recognition using surveillance cameras (Hao et al., 2023; Khan et al., 2020). Common approaches include identifying irises (Larregui et al., 2019), muzzle prints (Li et al., 2022a), and faces (Wang et al., 2020; Xu et al., 2021). However, acquiring such data often necessitates a controlled collection environment, such as deploying multiple cameras to capture images from various angles simultaneously (Hu et al., 2020) or using specialized cameras (Jaddoa et al., 2019). These

* Corresponding author at: Department of Computer Engineering, Mokpo National University, Muan 58554, South Korea.
** Corresponding author at: Department of Electronics Engineering, Jeonbuk National University, Jeonju 54896, South Korea.
E-mail addresses: shujiejulie@jbnu.ac.kr (S. Han), afuentes@jbnu.ac.kr (A. Fuentes), jbpark99@jbnu.ac.kr (J. Park), syoon@mokpo.ac.kr (S. Yoon), jcyang@tust.edu.cn (J. Yang), ycjeong@jbnu.ac.kr (Y. Jeong), dspark@jbnu.ac.kr (D.S. Park).

https://doi.org/10.1016/j.compag.2025.110301
Received 26 August 2024; Received in revised form 21 January 2025; Accepted 14 March 2025
Available online 19 March 2025
0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

requirements impose additional constraints on data collection and may disrupt normal cattle activities.

In most indoor farms, surveillance cameras are mounted at elevated positions to provide a comprehensive view of all cattle. In this setup, face recognition emerges as the most practical and universally applicable feature for identification. Consequently, cattle face recognition is the most suitable technology for real-world cattle monitoring in such environments.

A typical face recognition system consists of three fundamental components: face detection, face alignment, and face recognition (Wang and Deng, 2021). Mainstream research often focuses on these components individually to maximize performance improvements (Chen et al., 2022; Meng et al., 2023; Qiao et al., 2022; Yao et al., 2019). However, this approach has a notable drawback: it complicates the full automation of the recognition pipeline. For instance, in face recognition, state-of-the-art algorithms frequently rely on preprocessed cattle face images collected at specific times, training models to recognize these images. This paradigm typically requires specialized farm setups, such as facilities that restrict cattle movement (Jaddoa et al., 2019; Ruchay et al., 2023) or extensive manual preprocessing (Meng et al., 2024; Zhang et al., 2023). Data collected under these controlled conditions pose several limitations in practical applications. First, such datasets lack temporal characteristics. They usually consist of large volumes of data randomly divided into training, validation, and testing sets, often collected at specific times. As a result, even with a significant dataset size, the data remain homogeneous. This homogeneity leads to a decline in model performance when farm conditions deviate from the training set distribution. Second, these datasets typically feature a single cattle face per image, with the face occupying most of the frame. This setup poses challenges for real-world deployment, as cattle move freely on farms. Capturing constrained images of individual faces requires additional equipment, further complicating the process. These limitations hinder the algorithm's ability to automatically capture and recognize cattle faces in real-world conditions, presenting significant barriers to fully automating the entire recognition process.

Building on current research trends, we propose a new paradigm that removes constraints on cattle face collection, enabling simultaneous recognition of all cattle faces within a camera's field of view at any time of the day. To achieve this, we first construct a dataset with temporal characteristics, capturing cattle in free-range conditions. Next, we design an end-to-end cattle face recognition framework consisting of three components, incorporating farm-specific knowledge to reduce reliance on training data. This approach ensures robust model performance even when the environmental conditions during testing differ significantly from those in the training set.

The primary challenge in developing an end-to-end cattle face recognition framework lies in addressing the domain drift caused by the free-range activities of cattle. These factors introduce significant variations in the test dataset, deviating from the training dataset distribution and consequently affecting model performance after deployment. Frequently updating model parameters to mitigate these issues increases manual workload. Specifically, domain drift can result from variations in face poses caused by different barriers, changes in illumination due to weather, and shifts in local feature attention due to specific backgrounds. The limitations of the training datasets in capturing all potential variations further exacerbate these challenges, with temporal variations being a key driver of domain shifts in pose, illumination, and local feature attention over time.

Other advanced cattle face recognition studies have also explored the effects of pose and illumination variations. For example, Xu et al. (2024) proposed an algorithm that embeds diverse pose features through clustering, enhancing model robustness. Similarly, Ruchay et al. (2024) applied data preprocessing techniques to normalize image illumination before feeding data into the model, reducing the impact of lighting conditions. While these methods have made significant contributions, they are often constrained by upstream tasks such as face detection and alignment. As a result, their performance may degrade in fully automated end-to-end deployment scenarios.

To address the challenges in implementing cattle face recognition systems, we mitigate the impact of domain drift at the data level through knowledge and data-driven approaches. By combining farm knowledge with data augmentation techniques tailored to the cattle farming environment, we refine the training and inference processes of our cattle face recognition system. This knowledge includes temporal and spatial information of cattle faces, illumination changes, and environmental backgrounds, which are utilized as preprocessing and postprocessing references. Together, these factors enhance the training and inference processes, collectively improving the performance and adaptability of our system.

This paper presents three key contributions. First, we utilize Intersection over Union (IoU) matching of temporal targets to track cattle faces. Representative samples for recognition are selected from the tracked cattle face trajectories. Using facial landmark detection faces with five key landmarks are classified as frontal and prioritized as representatives. When frontal views are unavailable profiles are used. This strategy ensures the model consistently prioritizes test samples with a spatial distribution similar to the training set, improving recognition accuracy. Second, we mitigate the impact of illumination fluctuations on the model by augmenting the brightness, contrast, saturation, and hue in the training samples to cover diverse weather conditions. Although these augmentations are only applied during training, they significantly enhance the model's inference performance under varying lighting scenarios. Third, we minimize the bias of farm-specific data by excluding such information from the model input. By segmenting the foreground of the face and eliminating background obstacles, we enhance the model's integrity. This approach ensures the model concentrates solely on relevant facial features, regardless of farm-related influences.

To evaluate the effectiveness of our framework, we conducted experiments using a self-collected dataset comprising cattle face images captured over three days under varying weather conditions. Subsequently, we designated one day's data from this dataset for training and used the remaining for testing. Our experiments indicate that pose variation significantly impacts recognition accuracy, while illumination augmentation notably improves the model's generalization ability. Additionally, background removal enhances the model's reliability. Our model achieved a mean Average Precision (mAP) of 98.0%, confirming the feasibility of incorporating farm knowledge into a data-driven visual framework for cattle face recognition.

Integrating farm knowledge into the framework helps to achieve homogeneity in the distribution of training and testing data, thereby mitigating domain shift challenges in the test set. Furthermore, this approach reduces reliance on extensive data collection and minimizes labor requirements. Our method offers practical advantages for deploying precision livestock farming solutions, providing a streamlined implementation with reduced labor costs without compromising performance standards. The implementation has been made publicly available at https://github.com/shujiejulie/An-end-to-end-cattle-face-recognition-system.

2. Materials and methods

In this section, we provide a comprehensive description of our farm setup, data collection methodology, and proposed framework. First, we outline the farm environment, including key aspects such as the physical layout and prevailing conditions. Additionally, we elaborate on the camera setup employed, the data collection schedule, and the specific annotation guidelines for the collected data. Following this, we introduce the farm-specific knowledge and methodologies employed for face detection, alignment, and recognition experiments.

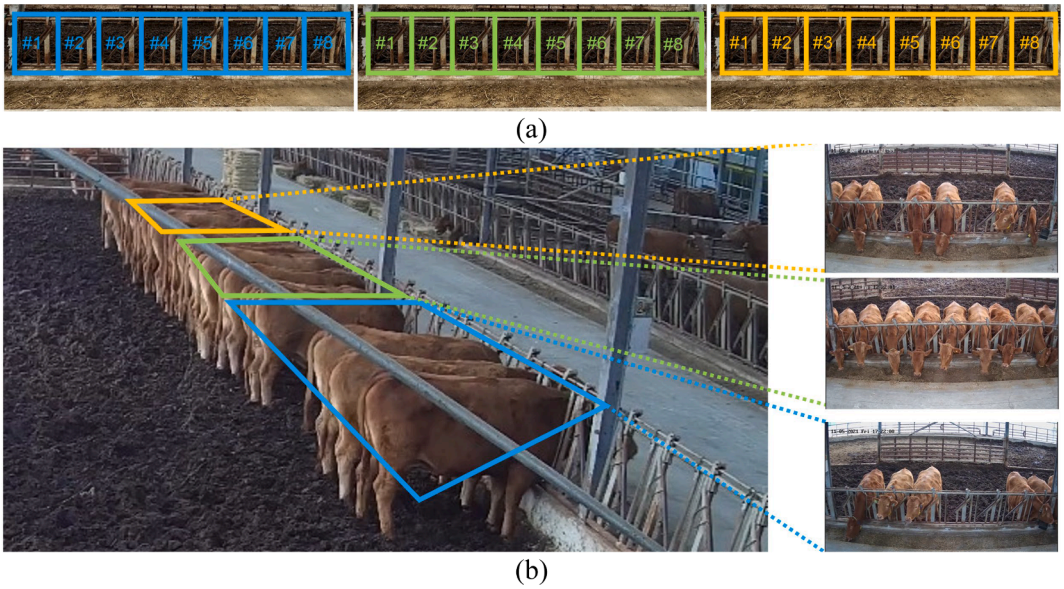


Fig. 1. Overview of the farm environment. (a) Mounting details of the surveillance camera with the barrier. (b) Perspectives from the three surveillance cameras.

2.1. Farm setup

To collect real-world cattle farm data using surveillance cameras, we positioned 3 Hikvision (HIK) surveillance camera devices with 4 K resolution (3840 × 2160) within an indoor farm measuring 30 m in length and 12 m in width (Meng et al., 2023). These cameras were strategically placed at an elevated height, facing the feed barrier, to comprehensively monitor cattle faces during feeding. To maximize the collection of face data, we conducted data collection during the cattle feeding period, as the cattle naturally protrude their heads through the barrier while eating, thereby facilitating the acquisition of video data with a substantial number of faces. Importantly, this approach does not restrict the movement of the cattle, as their feeding behavior occurs naturally.

As shown in Fig. 1**(a)**, the surveillance cameras were placed 6.6 m away from the feed barrier to avoid interfering with the cattle. Due to the field of view limitations of each camera, every camera can monitor only eight head spaces within the feed barrier. Each head space measured 0.83 m in length and 0.62 m in width. To achieve broader coverage, we deployed three cameras side-by-side, allowing us to monitor 24 consecutive head spaces in total. However, it should be emphasized that this coverage did not cover all head spaces within the barn, meaning that in some cases, some cattle were occasionally outside the camera's view.

The barn houses a total of 17 adult beef cattle and 2 calves. The calves were excluded from our study as they are managed under a different system compared to the adults. In contrast, the 17 adult cattle, as shown in Fig. 1**(b)**, share similar body shapes and fur colors. However, their facial characteristics exhibit significant variation, including differences in features such as horns, forehead shape, and fine-grained fur patterns. It is worth noting that these distinguishing features



Fig. 2. Cattle face annotations with bounding boxes and landmarks. The dataset used in the study includes various challenging weather conditions, including fog, sunshine, and cloud cover.

Table 1
Dataset Information. The values represent the appearances of cattle faces during the selected period, with a total of 3,000 images collected from each camera (1, 2, 3).

Cattle ID	Number of cattle face instances per day		
	November 5	November 6	November 9
Cattle#1	2959	2838	2652
Cattle#2	2000	2845	2794
Cattle#3	2874	2562	3000
Cattle#4	3000	3000	3000
Cattle#5	3405	3000	3000
Cattle#6	3000	3000	2974
Cattle#7	3000	2127	2941
Cattle#8	3000	2920	3000
Cattle#9	3000	3000	3000
Cattle#10	2923	3004	3000
Cattle#11	3000	1486	2000
Cattle#12	2948	3000	2949
Cattle#13	3465	2797	3000
Cattle#14	2930	2910	3000
Cattle#15	3303	2276	3000
Cattle#16	3000	2839	3000
Cattle#17	2966	3000	2000

extend beyond those explicitly mentioned.

2.2. Data acquisition

Following a similar farm setup, we collected two datasets from two farms. The primary dataset serves as the main resource for training and testing our model, with most of the experiments conducted on this dataset. Consequently, the primary dataset consists of a larger number of images. To ensure that our model not only maintains high performance on the primary dataset but is also adaptable to new farm environments, we introduced an additional dataset. Although smaller in scale, the additional dataset encompasses diverse scene variations and is exclusively used for fine-tuning to evaluate the model's scalability.

2.2.1. Primary dataset

To build our dataset, we collected surveillance video of cattle feeding sessions three times a day for three days (November 5, 6, and 9): morning, noon, and evening. While the precise feeding times varied in practice, they were approximately within the following ranges: 7:00–8:00, 11:00–12:00, and 17:00–18:00. During each of these periods, we annotated one minute of video data from each of the three cameras, capturing footage at 15 fps. This resulted in labeling 3,000 frames of data for each feeding session, with a total of 27,000 frames labeled, including bounding boxes and landmarks for the 17 cattle individuals, as shown in Fig. 2. Fig. 2 presents a subset of the labeled data collected across the three days, encompassing three different weather conditions: foggy, sunny, and cloudy. November 5 featured foggy weather, while November 6 was sunny, with sunlight reflections clearly visible on the cattle's bodies. On November 9, the weather was cloudy, with almost no sunlight observed.

Following previous works (Cao et al., 2019; Hewitt and Mahmoud, 2019), we annotated a total of 11 facial landmarks on the cattle, including the nose, mouth, eyes, bottom of the ears, top of the ears, and bottom of the horns. However, during model training, we used only five key landmarks: the nose, eyes, and the bottom of the ears. Our prior work (Han et al., 2022) demonstrated that the detection of horns was unreliable in some cases, and the mouth landmarks were prone to occlusion during feeding, which made the detection less consistent. Using these landmarks not only ensured the model's accuracy but also improved computational efficiency. Additionally, we annotated 300 images of cattle face masks captured on November 5. This dataset included a total of 1,700 entities which were used to train a semantic segmentation model designed to extract the foreground portion of the face accurately.

Table 1 provides a detailed summary of the dataset, where 3,000 frames per camera were captured daily. The dataset includes samples for 17 cattle, with each cattle contributing approximately 3,000 samples per day. Variations in sample counts occurred due to the free-ranging nature of the cattle. Some cattle were absent during the collection period, resulting in fewer than 3000 samples, while others moved along the boundaries between two cameras and were captured by both, leading to more than 3000 samples. Despite this, the sample counts for each cattle are generally balanced.

2.2.2. Additional dataset

We collected an additional dataset specifically designed to evaluate the scalability of our model. This dataset was obtained from a new farm and features 20 entirely new cattle, including 16 fully mature individuals and 4 juvenile individuals. The dataset offers greater environmental diversity, with data collected during different months and covering both daytime and nighttime conditions. The farm setup was also adjusted, with cameras installed at relatively higher positions. The dataset comprises a total of 2,800 images, of which only 800 were used to fine-tune the model, enabling it to adapt to the new farm's conditions. This dataset primarily serves to test the model's ability to generalize effectively across various farms and demonstrates its practical applicability. More details on the farm settings and dataset are elaborated in Section 3.5.2.

2.3. Methods

In this section, we introduce the farm-specific knowledge utilized to drive our algorithms, including spatiotemporal information, illumination, and background effects. We then describe the proposed framework, which encompasses algorithmic models for detection, alignment, and recognition. The detection phase primarily involves face detectors, while the alignment phase incorporates modules that apply farm-specific knowledge. In the recognition phase, we train the face recognition model using augmented samples.

2.3.1. Farm knowledge

We utilized three types of farm knowledge, spatiotemporal information, illumination variation, and background effects, to address three challenges in real-world cattle face recognition. Specifically, our temporal alignment module tracks samples belonging to the same cattle over time by utilizing temporal context. Since cattle are usually located near a fixed barrier during feeding, this facilitates temporal alignment. Based on spatial information, specifically the pose of the cattle's face, our pose alignment module can identify easily recognizable samples to represent the tracking trajectory for recognition. Temporal and spatial information are employed to address challenges caused by geometric variations in face pose, allowing for the consistent selection of samples with specific poses for recognition.

The utilization of illumination variation primarily relied on experimentation and expert knowledge. Images can be decomposed into illumination and reflection components (Du et al., 2023; Wu et al., 2020), where illumination causes variations in image brightness, and differential reflection between cattle and background causes contrast changes. These factors significantly impact inference performance. By augmenting training samples, the model has effectively addressed most of the challenges caused by illumination variations. Background effects were observed qualitatively during experiments. In data-driven modeling, the model's attention can sometimes shift to irrelevant features, such as barriers and feed, which are simple and easily distinguishable, thus diverting attention from facial features. Therefore, a foreground cropping module addresses this challenge. The following section describes the quantitative and qualitative aspects of the three types of farm knowledge applied in our work:

Spatiotemporal Information. The temporal context provides

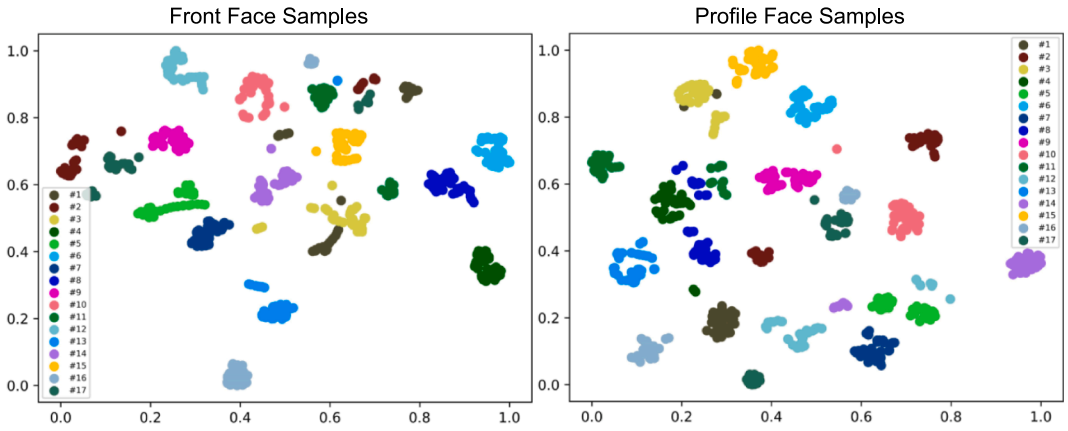


Fig. 3. Front and profile face samples used for t-SNE analysis.

Table 2
3D Landmark Templates for Cattle Faces.

Landmarks	Coordinates (mm)
Nose	(0, 0, 0)
Left eye	(45.0, 96.8, −96.8)
Right eye	(−45.0, 96.8, −96.8)
Left bottom of ear	(50.0, 106.1, −109.4)
Right bottom of ear	(−50.0, 106.1, −109.4)

information about the barrier where the cattle are located, which is relatively fixed and can be effectively utilized. However, spatial information varies significantly with the cattle's movements, so we focus on analyzing the spatial variations in the cattle's face. Cattle move freely within the barn, resulting in domain shifts in pose across different periods. We categorize cattle faces into fronts and profiles. Fronts are defined by the visibility of five landmarks: the nose, both eyes, and the bottom of both ears, whereas profiles do not display all these landmarks. Naturally, we find fronts easier to recognize than profiles.

Fig. 3 shows t-SNE analysis, derived from 45 samples randomly selected from 17 cattle, demonstrating that fronts cluster more effectively than profiles. Given the complexity of profiles, which may require separate consideration of left and right profiles, we chose to focus on frontal poses for reference knowledge. Frontal poses can be further refined using the Perspective-n-Point (PnP) algorithm to estimate rotation angles in 3D space. To support accurate estimation of facial rotation, we refined previous work on 3D cattle facial landmark templates by combining measurements from buffalo studies (de Melo et al., 2018) with field measurements on beef cattle. Table 2 presents the template coordinates of key facial landmarks relative to the nose, which serves as the origin (0, 0, 0), specifically listing the coordinates of the eyes and bottom of ears. These template coordinates provide a reliable reference for analyzing and comparing the spatial positioning of these landmarks across different cattle faces.

Fig. 4 **(a)** illustrates the cumulative distribution function (CDF) of rotation variations for all fronts in the training set. Faces can rotate in three directions in 3D space: pitch, yaw, and roll, as shown in Fig. 4 **(b)**. Specifically, rotations along the pitch axis are shown in blue, yaw axis in green, and roll axis in red. We estimated the rotation ranges for pitch, yaw, and roll to be $[-55, 55]$, $[-10, 89]$, and $[-65, 76]$ degrees, respectively. For convenience, these ranges were normalized to $[-1, 1]$. The rotation distribution of fronts in the training set is depicted in Fig. 5, indicating minor variations along the roll axis but significant variations along the pitch and yaw axes, aligning with natural behaviors. Pitch represents the up-and-down head tilt common during feeding, while yaw corresponds to lateral head movement, frequently observed in cattle at the periphery. Naturally, if the distribution of test fronts during inference aligns with this pattern, it can mitigate pose domain shifts during inference. The pose distribution of test fronts is discussed in Section 3.3.

Illumination Changes. The first row of Fig. 6 illustrates the significant variability in cattle face illumination across different times of the day and weather conditions, due to the diffuse reflection of fur and lighting conditions. As a cornerstone of our approach, we primarily focus on enhancing the brightness and contrast of training samples. The second row demonstrates that brightness enhancement can accommodate a wider range of lighting conditions. We randomly enhance training samples within a factor range of 0.5 to 1.5. The third row indicates that contrast enhancement can accentuate or diminish certain details, stemming from the disparate reflectance of the cattle's face and the background. We apply random enhancements within a factor range of 0 to 2.0. By superimposing varying lighting and contrast enhancements on our samples, our training dataset encompasses the most common Illumination changes, thus mitigating the impact of illumination domain drift during model inference.

Background effects. The barn environment, relatively uncomplicated, tends to facilitate the recognition of various backgrounds compared to facial features. Fig. 7 depicts attention maps generated by our model for face recognition. The first group (a)(b) illustrates instances where two distinct cattle are misidentified due to similarities in background features (e.g., forage shape). The second group (c)(d) demonstrates that upon background suppression, the model's attention shifts to facial identity information. In (c), the frontal view emphasizes the forehead and left eye of the cattle, while in (d), despite the less distinct forehead of the profile view, the model still attends to the left eye. These examples underscore how foreground extraction enhances the model's ability to generalize and interpret local feature attention shifts. Leveraging this farm-specific knowledge allows us to effectively mitigate background effects.

2.3.2. *Proposed framework*

The system framework is illustrated in Fig. 8. Fig. 8**(a) and 8(b)** show the models that require training within the system. Fig. 8**(a)** presents the face detector, which takes a complete surveillance image as input and outputs a mask, bounding box, and landmarks. We employed the single-stage YOLOv7-Pose (Wang et al., 2023) as the bounding box and landmark detector, and Mask RCNN (He et al., 2017) as the mask detector. Both detectors share the same input image, and their three types of outputs collectively operate on the same image.

Fig. 8**(b)** demonstrates the training process of the ID classifier. Based on the outputs of the face detector, the bounding box is used to automatically crop face samples from the input images, landmarks are stored as parameters for calculating cattle face pose variations, and the mask is used to exclude background regions from the cropped face samples. Illumination augmentation a data preprocessing technique, generates

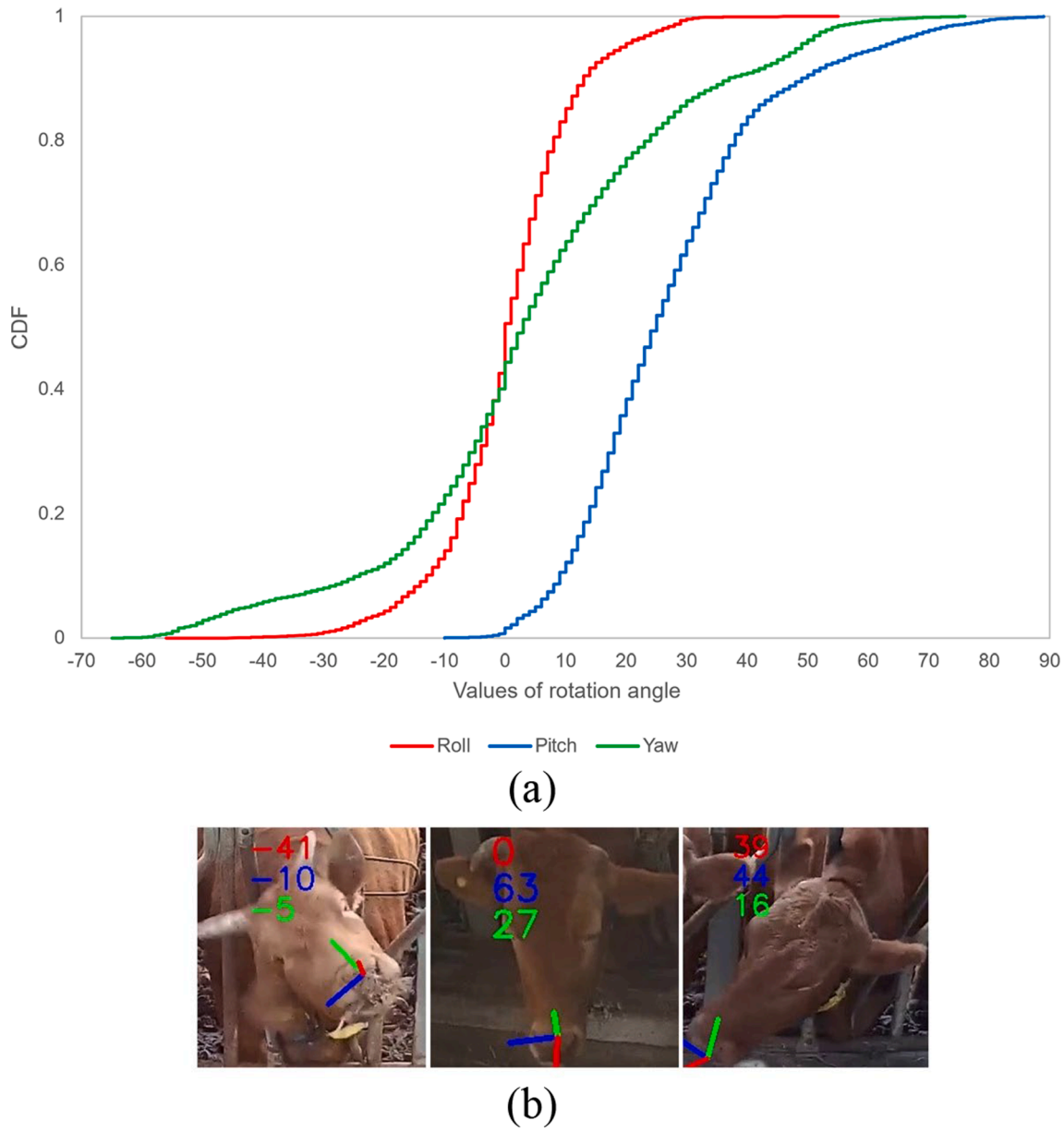


Fig. 4. Statistics of cattle face rotation. Blue represents pitch, green represents yaw, and red represents roll. (a) Cumulative distribution function of the rotation angles for all samples. (b) Examples of cattle face rotations in 3D space.

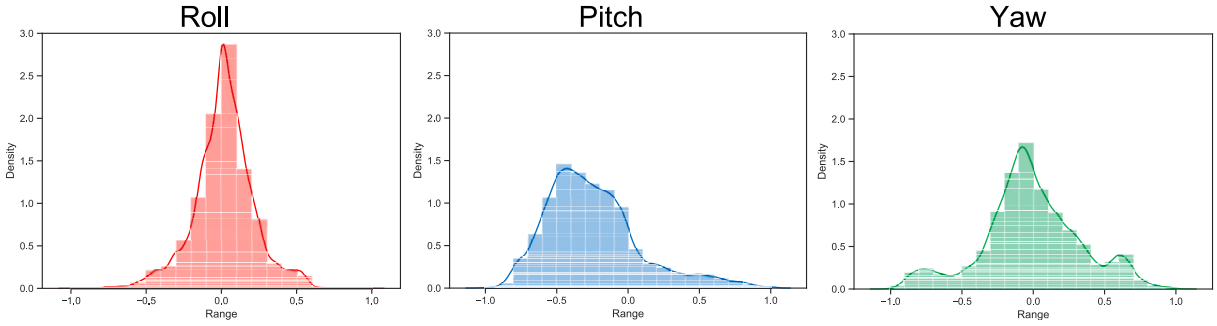


Fig. 5. Data distribution for the three rotational directions in the training set.

samples under varying lighting conditions offline, serving as training data for the classifier. We focused on enhancing the brightness and contrast of the samples, with specific parameter settings detailed in the experimental section. The ID classifier is a simple classification model consisting of a ResNet backbone and a fully connected layer for classification. Given the fixed number of cattle in our farm setup, the model's

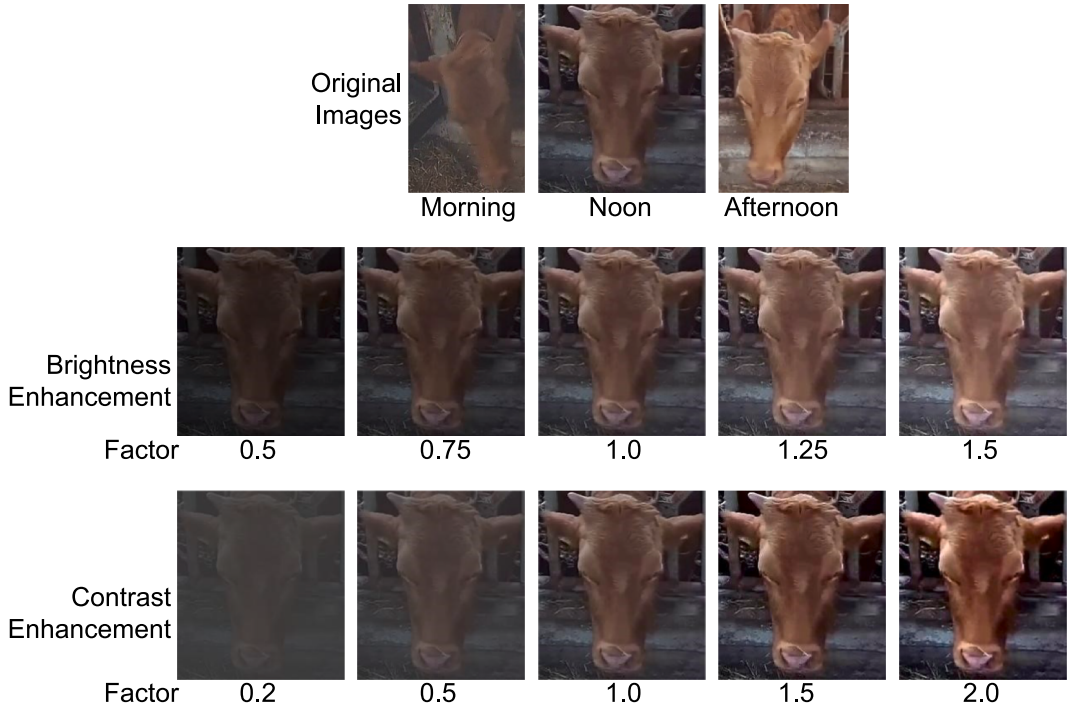


Fig. 6. Impact of illumination changes on training samples. Examples of the results after applying brightness and contrast enhancement.

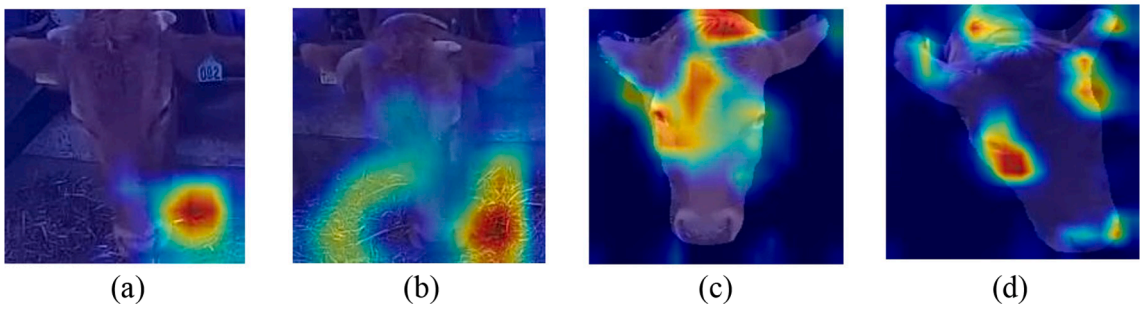


Fig. 7. Attention heat maps during cattle face recognition. In these maps, the model's attention is primarily focused on the background, particularly the area of forage with a similar shape.

output classes are directly set to 17. The ID classifier takes preprocessed cattle face samples as input and outputs the cattle ID corresponding to the highest confidence score.

Fig. 8**(c)** depicts the workflow of the model during inference. The three outputs of the detector are used for foreground cropping, temporal alignment, and pose alignment, respectively. Foreground cropping masks the background and extracts facial samples based on the detected foreground region. Temporal alignment tracks each cattle temporally using the IoU between bounding boxes of consecutive frames. The results of cropping and tracking are aggregated into trajectories, containing different facial samples of the same cattle over time. Pose alignment classifies samples into fronts and profiles, selecting the more easily recognizable samples to represent the entire trajectory as input to the ID classifier for ID prediction.

Fig. 8**(d)** illustrates temporal alignment, where the blue bounding box represents the detection result at time $t = 0$, and the yellow bounding box represents the result at $t = 1$. Tracking is achieved based on the IoU between these positions. Inspired by a tracking algorithm called SORT (Bewley et al., 2016), we match cattle faces across consecutive frames by comparing their positions and associating them as part of the same trajectory. Tracking cattle faces leverages spatial characteristics to assign temporal properties to detection results. This approach reduces the model's burden by eliminating the need to recognize all faces within a trajectory, requiring only the recognition of the "easier" ones.

Fig. 8**(e)** demonstrates the pose alignment workflow, where fronts are prioritized as representative inputs to the classifier. Based on the number of detected landmarks, samples with all five landmarks detected are classified as fronts, while others are categorized as profiles. In cases where fronts are unavailable or have low confidence, profiles are used instead.

Specifically, during the training phase, the detector is trained to detect the faces of cattle only when they extend their heads through the barriers to eat, while all instances of cattle moving around the barn are used as negative samples. Foreground cropping isolates the primary objects of interest and extracts face samples by performing bitwise operations between the detected masks and the original images, effectively removing irrelevant background elements. Illumination augmentation then generates samples with varying brightness, contrast, saturation, and hue to support the training of the ID classifier. It is important to note that this data includes all fronts and profiles. During training, we aim for the model to retain the ability to recognize various pose variations, enabling it to handle extreme situations, such as when no fronts are available throughout an entire period. These processes significantly

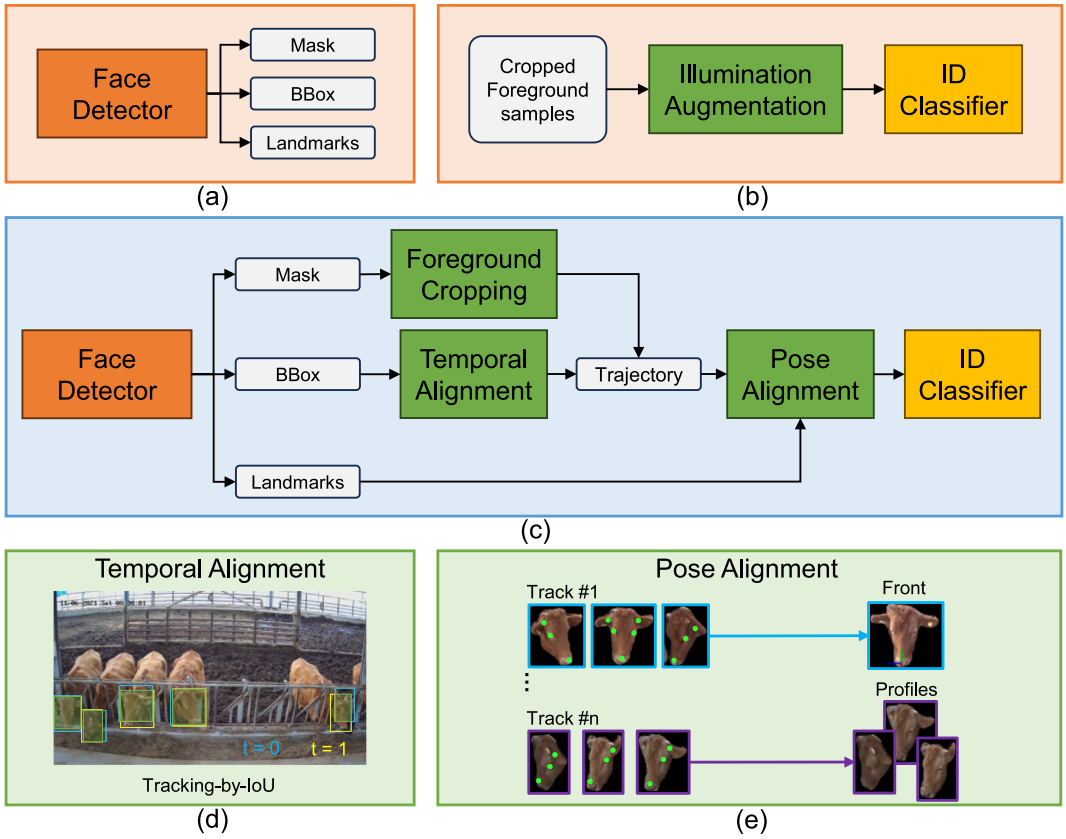


Fig. 8. Overview of the proposed framework. (a) and (b) illustrate the two modules that require training: the detector and the ID classifier. (c) presents the system's inference workflow, where the detector's results are processed by three knowledge-driven modules: foreground cropping, temporal alignment, and pose alignment. The aligned representative samples are then fed into the classifier for ID recognition. (d) details the temporal alignment process, which is based on an IoU tracker. (e) depicts the pose alignment process, selecting representative samples from each tracked trajectory for input into the ID classifier.

enhance the accuracy and reliability of our model's predictions, making it well-suited for practical applications in complex and dynamic environments.

During the inference process, unlike the random inputs used during training, the input video sequences possess temporal characteristics. Temporal alignment tracks the same face across frames by linking samples with a bounding box IoU greater than 0.5 between consecutive frames. All faces in a video are tracked frame by frame, and the results are assigned temporary IDs. Pose alignment selects the fronts from each trajectory as the representative for the entire trajectory. The identity with the highest confidence score above 0.99 (Top-1) is assigned as the ID for the entire trajectory. If there are no fronts or if the confidence scores for the fronts are low, profiles are used instead, with the same requirement of a confidence score above 0.99. If no samples meet this criterion, the trajectory is marked as unrecognized ($ID = -1$).

3. Results

This section details the experimental setup, detection results, pose alignment results, and recognition results of the system, followed by an ablation study to compare the contributions of each module. The experimental setup details the deep learning models and training strategies selected for our main modules. The detection results are derived from an independently trained detector, which provides ample samples for subsequent detection. The distribution of fronts rotation confirms the reliability of our standard, using five visible landmarks to define a front. The recognition results represent the system's final outcomes, encompassing all farm knowledge and data-driven modules. We use a baseline model that does not incorporate farm knowledge and directly identify all cattle faces in each frame. To highlight the importance of each module, we evaluate the performance of modules in the ablation study. Temporal and pose alignment enhances recognition performance directly during inference, while foreground cropping and illumination augmentation require model parameter training.

3.1. Experimental setting

To enable long-term cattle recognition, it is important to consider that environmental variations in the data across different days and times may influence the model's performance. To assess the impact of these environmental changes on our dataset, we adopted a specific approach. Data from November 5 were designated as the training set and validation set for training all system modules, while data from November 6 and November 9 were used as testing sets. The training set consisted of 7,200 frames captured during the morning, noon, and afternoon, with 1,800 frames allocated as the validation set. Additionally, 9,000 frames from each of the two test days were set aside for testing.

This unconventional data partitioning approach was designed to demonstrate that our method does not rely heavily on the training data. By adopting a knowledge- and data-driven strategy, the model leveraged farm knowledge to identify samples with temporal invariance from a test dataset that was twice the size of the training set. These samples, which shared a similar distribution with the training set, were accurately identified, showcasing the practicality and scalability of our method.

To implement the detection function within the system, we employed two models. YOLOv7-Pose was used to detect the bounding box and landmarks of faces, trained on the aforementioned 7,200 frames. Image augmentations such as rotation, translation, scaling, shearing, and flipping were applied during training as preprocessing steps to enhance detection stability. These techniques enabled the model

Table 3
Cattle Face and Landmark Detection Results.

Detection		mAP50 (%)	mAP50:95 (%)	mAR50:95 (%)
Face	Faster R-CNN (Ren et al., 2015)	98.9	77.0	81.9
	FreeAnchor (Zhang et al., 2019)	98.6	76.8	83.0
	Dab-DETR (Liu et al., 2022)	98.9	75.5	82.2
	DINO (Zhang et al., 2022)	98.9	79.8	86.4
	YOLOv7-Pose	98.7	78.2	96.6
Landmarks	CPM (Wei et al., 2016)	96.6	51.1	61.3
	IPR (Sun et al., 2018)	95.2	44.6	61.4
	LiteHRNet (Yu et al., 2021)	96.8	53.0	61.3
	SimCC (Li et al., 2022b)	96.7	50.2	61.1
	YOLOv7-Pose	97.0	52.4	67.0

to detect faces in various poses, generating sufficient samples for subsequent modules. Mask-RCNN was used to extract the foreground region of faces, with a pre-trained model fine-tuned using 300 face mask images. Both detectors were trained using the SGD optimizer, with learning rates set to 0.00001 and 0.0025, respectively.

The ID classifier serves as the recognition module, comprising a ResNet100 backbone (He et al., 2016) and a Partial FC classification layer (An et al., 2022). The training data consists of 45,229 face samples derived from the detector's results, covering 17 categories. Illumination augmentation techniques—including random adjustments to brightness, contrast, saturation, and hue—were applied with random seeds of 0.5, 1.0, 0.2, and 0.1, respectively. The classifier was trained using the ADAM optimizer with a learning rate of 0.001.

3.2. Detection results

Table 3 showcases the performance of the detection algorithm in precisely detecting cattle faces within our dataset. Face detection performance is evaluated based on Intersection over Union (IoU) precision, while landmark detection is assessed using Object Keypoint Similarity (OKS). In this evaluation, the model's prediction is considered correct if the IoU or OKS between the prediction and ground truth meets or exceeds 0.5. To compute the mAP at IoU threshold 0.5–0.95 (mAP50:95), we calculated the average precision across all results ranging from IoU or OKS thresholds of 0.5 to 0.95.

As shown in Table 3, face detection and landmark detection are typically performed by separate models. However, YOLOv7-Pose is capable of detecting both simultaneously, significantly reducing the computational cost of model deployment. Moreover, YOLOv7-Pose achieves accuracy comparable to other state-of-the-art detection models, achieving the highest mAP50 of 97 % in landmark detection.

While the precision slightly decreased with a higher threshold, face detection remained above 78%, and landmark similarity showed greater sensitivity to threshold variations compared to face detection. For the cattle face alignment task, the primary role of landmark detection was to differentiate between fronts and profile views. Thus, using a threshold of 0.5 to assess the landmark predictions proved to be sufficient for this purpose. Our algorithm consistently delivered promising results, accurately detecting nearly all cattle faces regardless of their pose, providing a substantial number of samples for downstream tasks.

Fig. 9 illustrates the results from the detection model, which directly outputs bounding boxes and landmarks for cattle faces. Ground truth annotations are shown in green, while the model's predictions are marked in orange. The confidence scores of the predictions are indicated by orange numbers. Observations reveal that predictions with confidence scores of 0.9 closely align with the ground truth annotations. This demonstrates that the model's high-confidence predictions accurately correspond to the actual cattle face locations and landmarks, indicating accurate performance.

3.3. Pose alignment

Initially, we assessed the feasibility of pose alignment by calculating the distribution of front rotations using ground truth landmarks from the training set. Similarly, we estimated the distribution of front rotations in the test set based on the detector's output. The results are shown in Fig. 10.

The frequency distribution histograms of the samples of the training set (November 5) and the test sets (November 6 and 9) across the three rotational directions indicate that both the training and test sets exhibit similar distributions. This demonstrates the viability of using five landmarks as a criterion for selecting frontal faces is feasible, allowing our model to operate within the same data domain. Consequently, this confirms that the training set positively influences system inference.

3.4. Recognition results

In this section, we validate the training results of the ID classifier. Fig. 11 presents the training and validation accuracy, along with the loss curves, for the ID classifier. The ID classifier is trained on cropped face samples, with the output being their corresponding IDs. As shown in Fig. 11, the ID classifier achieves 100 % accuracy on the training set, but the highest accuracy on the validation set is just above 80 %. This indicates a significant difference in data distribution across different time periods in our dataset, though the model does not exhibit overfitting.

It is important to note that the cattle face recognition system does not produce a single cropped face as its final result. Instead, it generates multiple results, including face IDs, locations, and landmarks derived from detection. To comprehensively evaluate the performance of the complete end-to-end framework, we use both classification and detection metrics. For classification metrics, we directly calculate the Average Precision (AP) for each category during the corresponding period and plot the Receiver Operating Characteristic (ROC) curve for each category. For detection metrics, we use the COCO metric to evaluate the

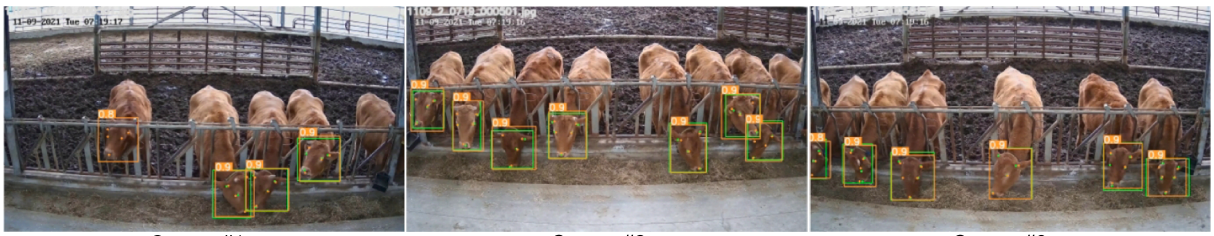


Fig. 9. Example output of YOLOv7-Pose. The output includes bounding boxes and landmarks, where green indicates the ground truth and orange represents the predictions.

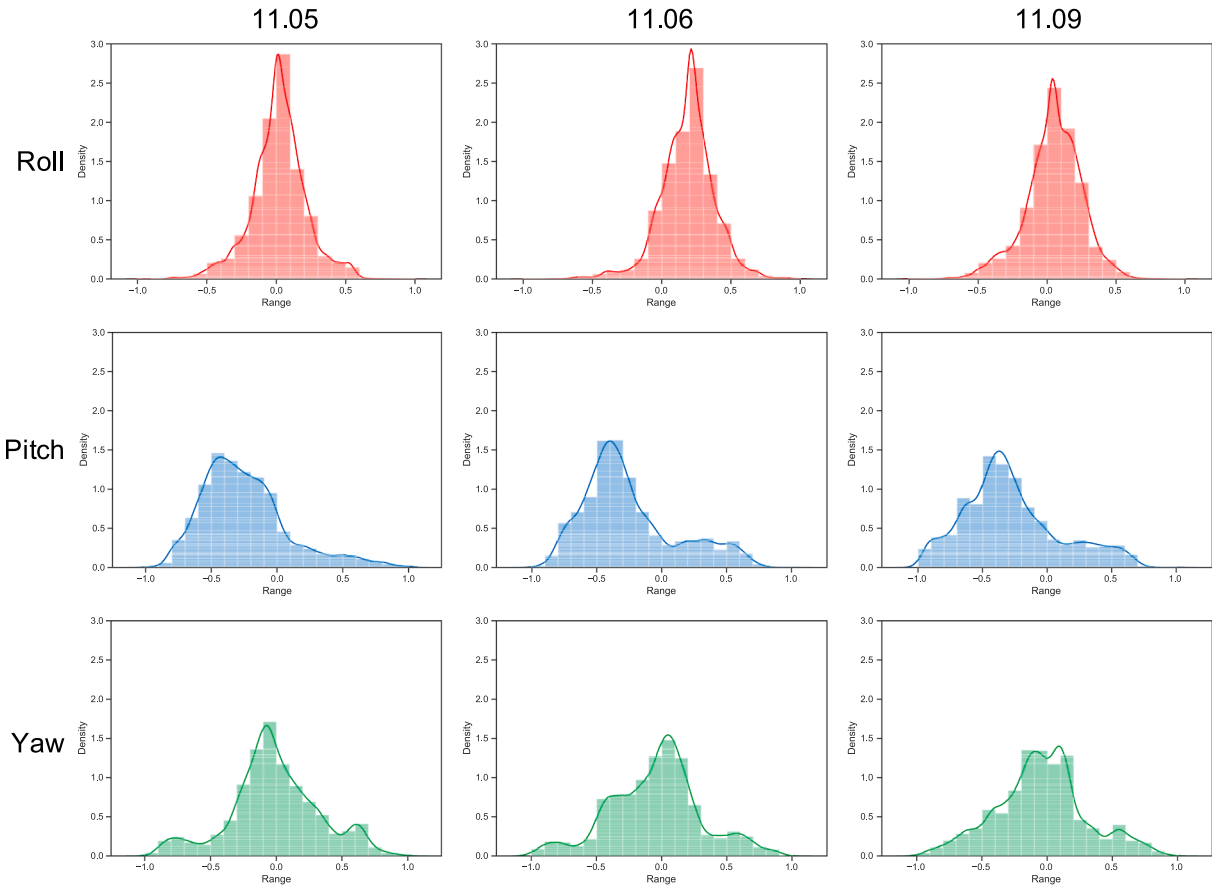


Fig. 10. Data distribution across three rotational directions in the training and test sets. The training set is represented by data from November 5 (11.05), while the test sets are from November 6 (11.06) and November 9 (11.09). The figure shows that after screening for pose rotation scope, the distribution of test set data closely aligns with the training set across all rotation directions.

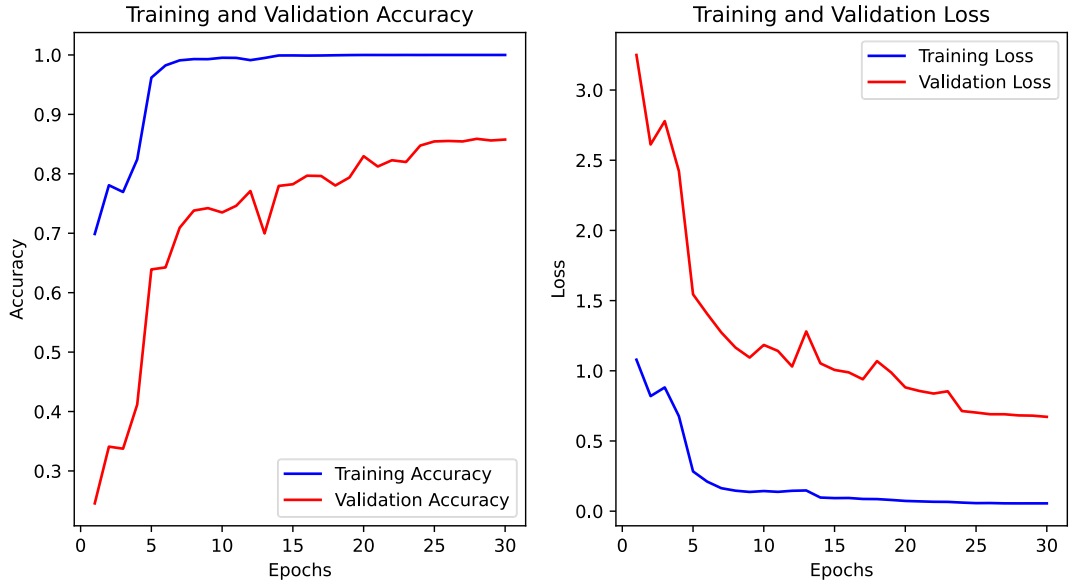


Fig. 11. Performance analysis of the ID classifier. Training and validation accuracy and loss are tracked over 30 epochs.

model's performance and analyze the impact of temporal domain shift on the model across different periods.

Table 4 presents the recognition AP for each category over the corresponding periods, showing that the majority of cattle were recognized with high accuracy, with AP values approximating 99 %. However, there were few exceptional cases: *cattle#11* did not engage in feeding during the periods 11.06-morning and 11.09-noon, leading to recognition results of *None*. A similar situation was observed for *cattle#17* during 11.09-afternoon.

Only a few cattle showed AP values below 95%, including *cattle#5*

Table 4
AP results for each cattle in each time period.

Monitoring time	Cattle IDs																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
11.06-morning	1.0	1.0	1.0	1.0	0.917	1.0	1.0	0.999	1.0	0.999	None	0.98	1.30	0.998	1.0	0.996	1.0
11.06-noon	0.996	0.988	0.995	0.999	1.0	1.0	1.0	1.0	1.0	0.818	0.5	0.999	1.0	1.0	0.996	0.971	1.0
11.06-afternoon	1.0	1.0	1.0	1.0	1.0	1.0	0.995	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.94
11.09-morning	0.999	1.0	1.0	1.0	1.0	0.986	1.0	1.0	1.0	0.885	1.0	1.0	0.997	1.0	1.0	1.0	1.0
11.09-noon	0.999	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	None	0.855	0.851	1.0	1.0	1.0	1.0
11.09-afternoon	0.997	0.994	1.0	1.0	1.0	0.856	1.0	1.0	1.0	0.999	0.658	1.0	1.0	1.0	0.996	1.0	None

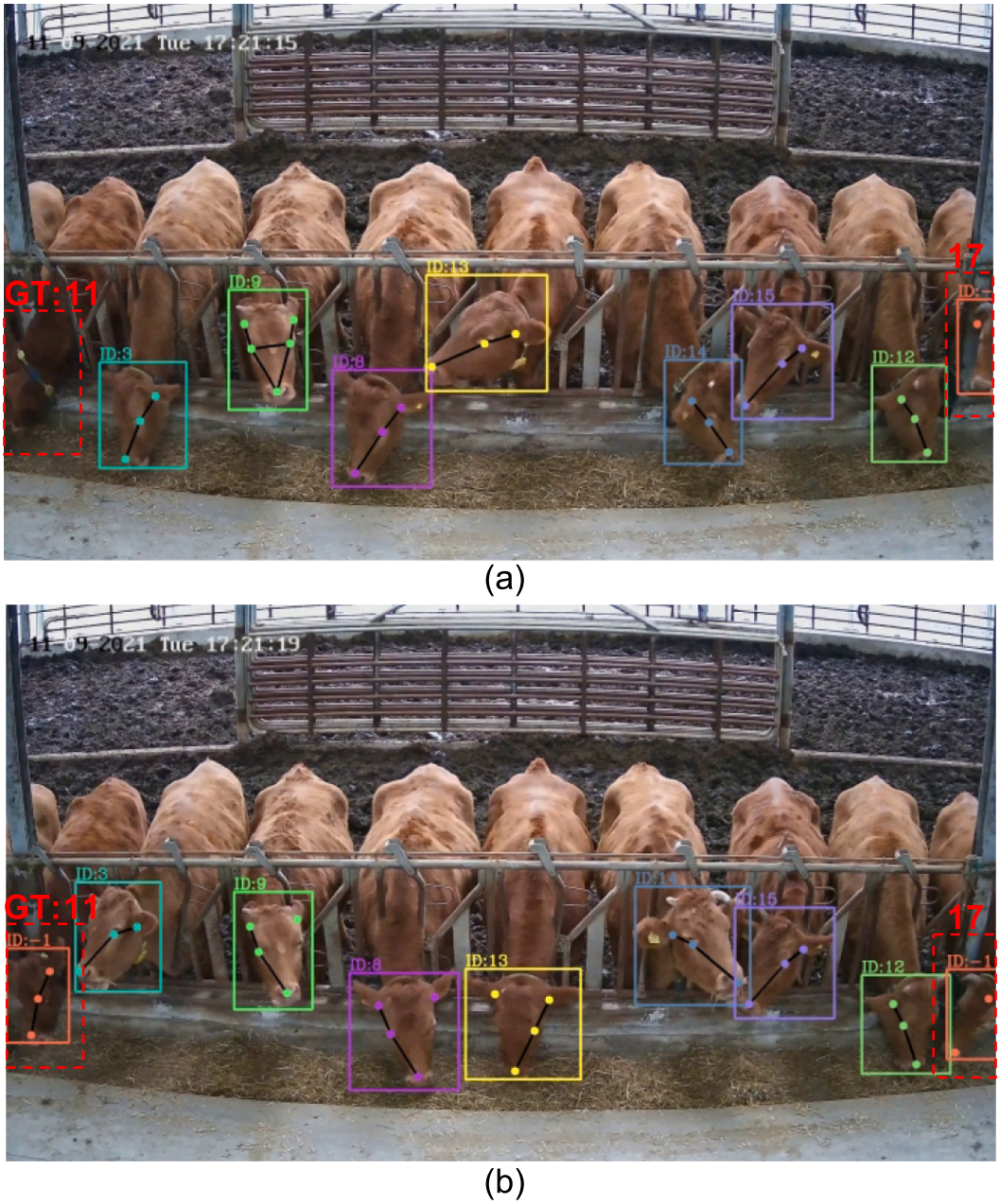


Fig. 12. Demonstration of unsuccessful cases in the cattle face recognition system, with dashed boxes representing ground truth and solid boxes showing prediction results. (a) shows cases of failed detection, while (b) illustrates incorrect recognition results.

on 11.06-morning, *cattle#6* on 11.09-afternoon, *cattle#10* on 11.09-morning, *cattle#10* and *#11* on 11.06-noon, *cattle#11* on 11.09-afternoon, *cattle#12* and *#13* on 11.09-noon, and *cattle#17* on 11.06-afternoon during specific periods. Two primary reasons account for this phenomenon. First, some cattle appeared at the edges of the camera's field of view, causing them to move their faces out of sight while feeding, resulting in missing facial features, similar to occlusion. Second, a few cattle exhibited significant movement while feeding, potentially

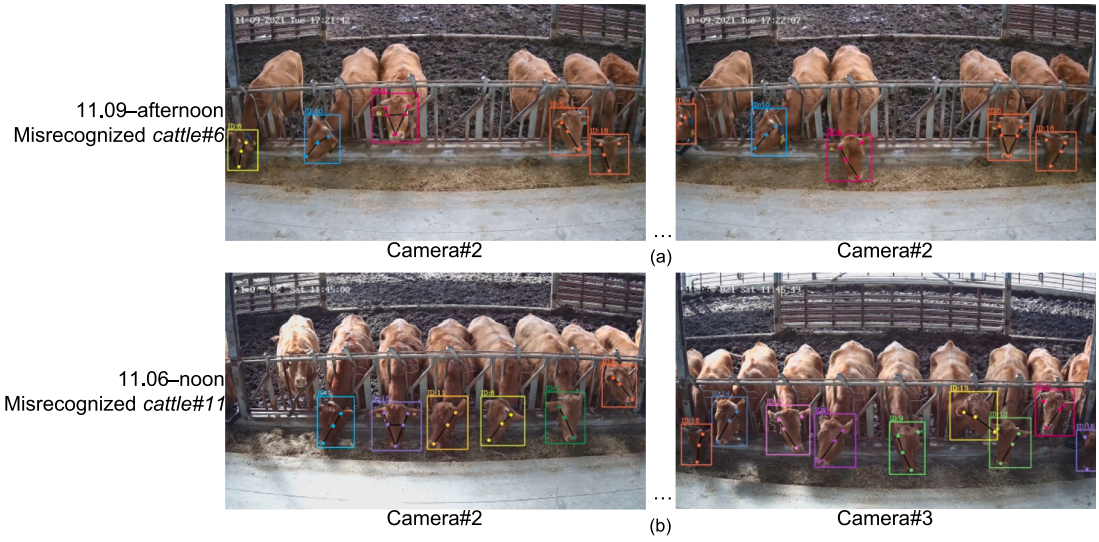


Fig. 13. The system's recognition results are illustrated with two examples. (a) shows the recognition results within the same camera's field of view during the same time period, while (b) presents the results across different camera fields of view during the same time period.

crossing between camera fields. For example, *cattle#11* had an AP of only 0.5 during 11.06-noon, due to this behavior. As it moved from camera#2 to camera#3, it was misrecognized in camera#3 but accurately recognized in camera#2, resulting in an average AP of 0.5. These cattle face images contained some occluded samples due to the limitations of the camera's field of view, which represents a special case not covered by our recognition system. However, overall, these special samples did not significantly impact the model's performance, as samples with severe occlusion were identified as -1.

Fig. 12 presents examples of typical misrecognitions, corresponding to the low recognition performance of *cattle#11* on 11.09-afternoon as shown in Table 4. In the figure, red dashed lines indicate the misidentified cattle and their corresponding ground truth IDs. In Fig. 12**(a)**, the ground truth for *cattle#11* was not detected, resulting in failure to recognize it. Meanwhile, *cattle#17*, which exposed only a very small portion of its face, was detected. However, such samples cannot be correctly identified by the system as they lack sufficient facial features for reliable recognition. These errors stem from inaccuracies in the detector. Fig. 12**(b)** shows cases where two cattle were assigned an ID of -1. This occurred because, although the detector successfully identified the cattle, the detected samples could not be assigned to an ID by the ID classifier with high confidence, leading to their classification as -1. These errors are due to inaccuracies in the ID classifier. In summary, all such errors occurred at the edges of the camera's field of view, where significant loss of facial features led to these misidentifications.

Some qualitative recognition examples are also shown in Fig. 13, illustrating instances of misidentified cattle during the same period. (a) demonstrates the misrecognition of *cattle#6* within the same camera's field of view. Due to its position at the edge of the field, significant head movements cause a large portion of its face to leave the camera's view. This leads to an interruption in the tracking trajectory, and when the face reappears, it is tracked as a new trajectory. However, the samples within this new trajectory are heavily occluded, leading to misrecognition. (b) illustrates another scenario where the interruption of tracking due to cross-camera movement results in recognition errors. *Cattle#11* moves from the field of view of camera#2 to the edge of camera#3. For similar reasons, the samples of *cattle#11* in camera#3′s view are occluded and misrecognized as *#15*. Due to the lack of cross-camera synchronization in our system, the misrecognition of *cattle#15* in camera#3 was still accepted by the system, despite the correct identification of *#15* in camera#2.

The ROC curve in Fig. 14 shows similar results, calculated by

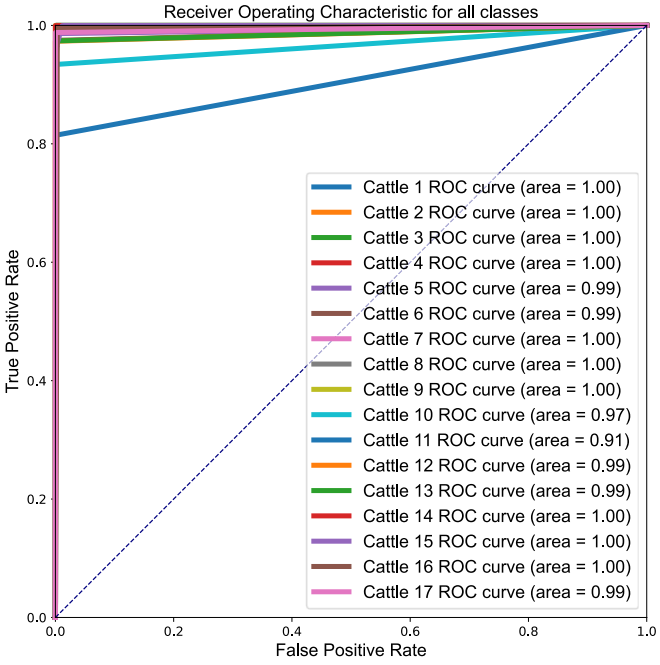


Fig. 14. The ROC curves for the ID classification results of each cattle, where a curve closer to 1.0 indicates better performance.

Table 5
Cattle face recognition involves providing quantitative results for specific periods as well as an overall average.

Models	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	mAR (%)
Baseline	55.13	53.76	42.04	44.51
VGGFACE2(Ruchay et al., 2024)	62.28	55.95	44.99	48.20
MobileFaceNet (Xu et al., 2024)	63.12	57.44	45.84	50.27
Ours-top-n	94.72	87.54	69.12	73.79
Ours-top-1	98.0	90.01	71.17	75.3

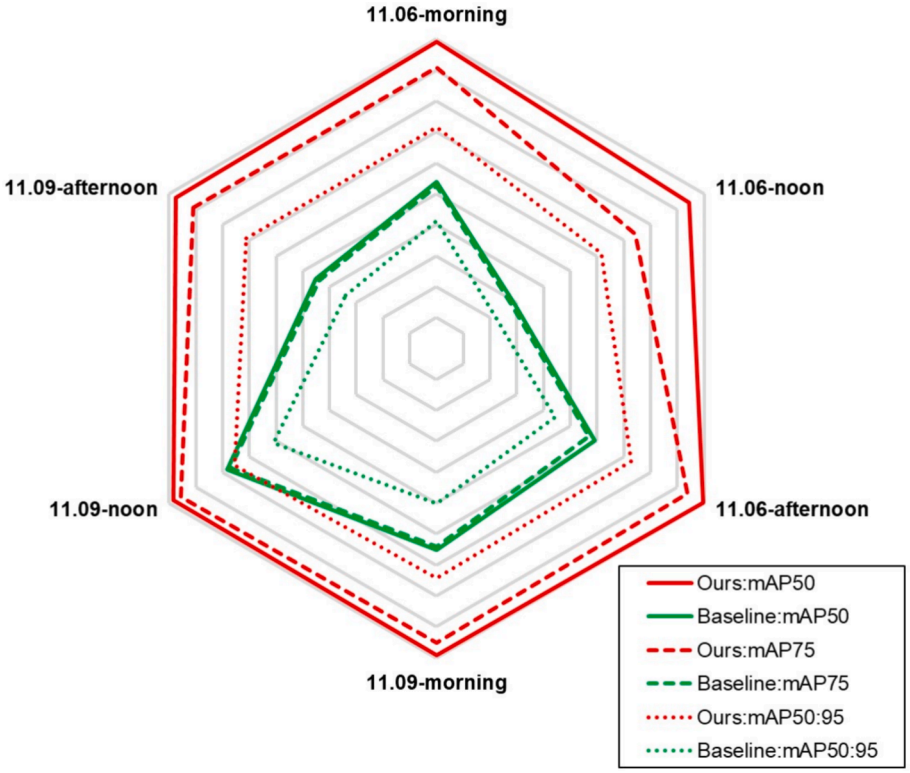


Fig. 15. Radar charts of COCO evaluation metrics for each time period.

aggregating the recognition results for all periods over two days. Cattle at the edges of the camera's field of view, such as *cattle#10* and *#11*, showed performance degradation, while cattle in the middle of the field of view generally exhibited good performance with an area under the curve close to 1.

To highlight the model's ability to effectively detect faces, we evaluated it using mAP and mean Average Recall (mAR) object detection metrics. The quantitative results are shown in Table 5, presenting the average outcomes for all video inputs for all models. The baseline represents the workflow without any knowledge-utilizing modules, using the ID classifier to directly recognize cattle face IDs without foreground cropping, illumination augmentation, or alignment.

Our method (ours) has two versions, with the primary difference being the selection of representative samples in the trajectory: ours-top-1 selects the single representative sample with the highest confidence, assigning its recognized ID to the entire trajectory, while ours-top-n selects the most frequently recognized ID among all representative samples, with n indicating the number of representative samples in the trajectory, which varies dynamically.

Additionally, our method was compared with other state-of-the-art cattle face recognition models. We selected two recent models, VGGFace2 and MobileFaceNet, as they were pretrained on human face datasets. However, direct performance comparisons are challenging due to differences in the cattle face dataset and the upstream tasks, including detection, alignment, and preprocessing. To ensure a fair comparison, we replaced the ID classifier in our framework with the VGGFace2 and MobileFaceNet models, while keeping the outputs of the upstream tasks consistent.

As a result, the baseline, VGGFace2, and MobileFaceNet rely solely on data-driven approaches, which makes their performance suboptimal when the test and training datasets exhibit different distributions. Nevertheless, compared to the baseline, VGGFace2 and MobileFaceNet show noticeable improvements, demonstrating the benefits of pretrained models and targeted preprocessing.

Compared to the baseline, ours-top-1 improved performance by 42.87% in mAP50, with at least a 29.13% enhancement in other metrics. This improvement is attributed to the knowledge-driven contributions. Foreground cropping and illumination augmentation ensure the trained model can handle various scene changes and lighting conditions. Tracking facilitates the selection of more easily recognizable samples, which serve as the ID for the entire trajectory. The benefits are evident: directly recognizing all samples with the model would lead to errors due to the presence of more challenging samples. Detailed comparative results across different times and dates are shown in Fig. 15, demonstrating that ours-top-1 significantly outperforms the baseline. The radar chart for the baseline reveals that data from November 9 is relatively easier to recognize than that from November 6, with the highest accuracy achieved during 11.09-noon but poor performance during 11.06-noon. This indicates the baseline's instability over time, a problem our model effectively addresses by showing consistent performance across six different periods, including the challenging 11.06-noon. The mAP50 radar chart for our model approximates a regular hexagon. From this analysis, it is evident that leveraging temporal and pose alignment can greatly enhance model performance without the need for additional parameter training.

3.5. Ablation study

In the ablation study, we evaluated the effectiveness of our model from two perspectives. First, Module Analysis was conducted to compare the contribution of each module to the model. The results demonstrated that all modules significantly enhanced the model's performance. Second, the Generalization Study was carried out to assess the generalizability of our model across different farms. The experiments showed that with minimal fine-tuning, our model could effectively adapt to new entities in a new farm environment.

3.5.1. Module analysis

The ablation study aims to evaluate the impact of each module on the system workflow, including foreground cropping (FC), illumination

Table 6
Performance Results of Different Modules in the System.

Models	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	mAR (%)	fps
BL	55.13	53.76	42.04	44.51	13
BL + FC	63.55	62.08	48.34	51.02	8
BL + FC + IA	82.17	78.82	61.34	66.17	8
BL + FC + IA + TA	97.67	89.66	70.84	74.94	10
BL + FC + IA + TA + FF	90.95	83.57	66.07	69.84	13
BL + FC + IA + TA + FF + PF	98.0	90.01	71.17	75.3	13

augmentation (IA), temporal alignment (TA), and pose alignment. It is important to note that there are two different scenarios for pose alignment: First, using only selected fronts as trajectory representatives (denoted as FF). Second, using profiles (denoted as PF) when no fronts are selected or when their confidence is below 0.99. Among these modules, FC corresponds to the model's ability to mask cluttered backgrounds, IA corresponds to its ability to handle illumination variations, and the combination of TA, FF, and PF corresponds to its ability to manage pose variations.

Table 6 shows that FC slightly improves model performance but increases system runtime due to the additional computational overhead from mask detection. Additionally, incorporating the IA module significantly enhances performance, indicating that illumination variation over time and domain shifts are major factors affecting model performance. Applying TA results in a further significant performance boost since all faces in the trajectory are recognized, and the highest confidence sample is selected. Compared to not using TA, the system's fps is increased to 10. Without TA, the ID classifier can only identify samples frame by frame. However, with TA, the samples from all frames are aggregated before inputting them into the classifier, reducing the number of classifier invocations and thus improving the runtime. In contrast, applying TA + FF accelerates the computation speed to 13 fps, similar to BL, but leads to performance degradation. This decline in performance is due to a significant portion of the samples being ignored, as some cattle might not show their front faces for extended periods. When using TA + FF + PF, system performance reaches a peak of 98.0% while maintaining a speed of 13 fps. This confirms that prioritizing fronts enhances accuracy, while supplementing with profiles increases robustness. Furthermore, because TA + FF + PF sets priorities, the model operates faster.

Fig. 16 illustrates a comparative example of using TA + FF + PF versus not using it. The x-axis represents the video frame sequence (1 to 1000), while the y-axis represents the recognized cattle ID (1 to 17). With TA + FF + PF or TA + FF, the model consistently recognizes the target *cattle#9* throughout the video sequence. Without any alignment (BL + FC + IA), the model makes multiple errors throughout the video, often misidentifying the target as *#13*. Therefore, using the top-1 accuracy of representative samples as the recognition standard proves beneficial.

3.5.2. Generalization study

To evaluate the generalizability of our method across different farms, we tested it on the additional dataset collected from another farm. This dataset primarily enhances the diversity of the data. It includes four barns and a total of 20 cattle. The data collection spans a wider range of time, covering July, August, September, and December. The dataset

Table 7
Performance Results of the Additional Dataset. Characteristics represent the distinct variations in each segment of data.

Characteristics	mAP50 (%)	mAP75 (%)	mAP50:95 (%)	mAR (%)
July Morning – Adult Individuals	81.5	27.06	37.05	42.89
August Morning – Juvenile Individuals	73.58	40.19	40.87	45.15
September Afternoon – Adult Individuals	99.43	45.86	51.69	59.28
December Evening – Adult Individuals	94.52	48.34	50.92	56.6

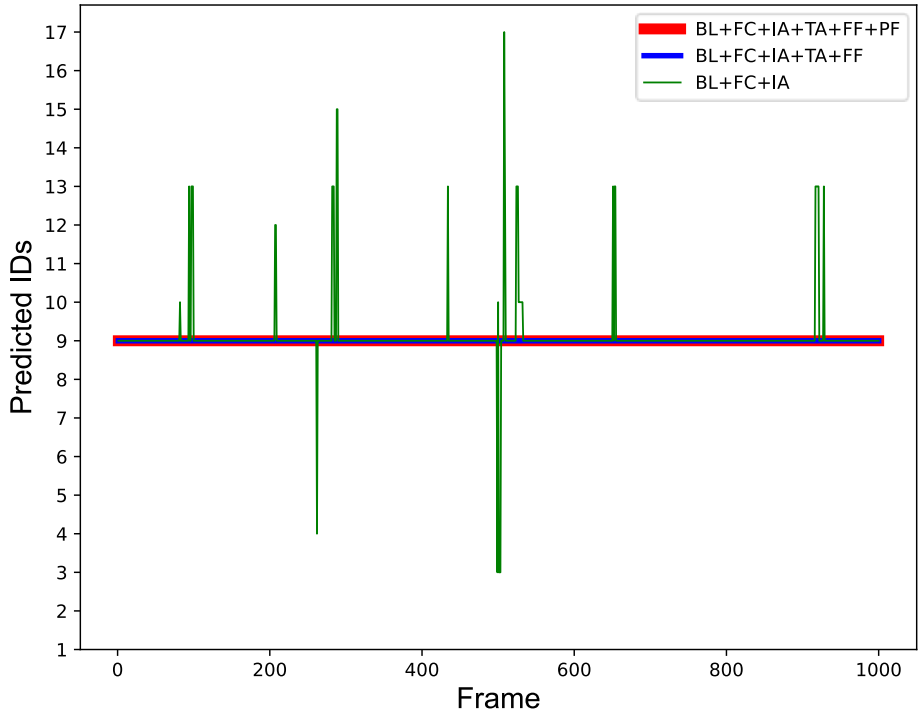


Fig. 16. An example illustrating the role of temporal and pose alignment during inference. The plot shows the recognition results for cattle #9, with the x-axis representing video frame numbers and the y-axis representing the predicted cattle ID.

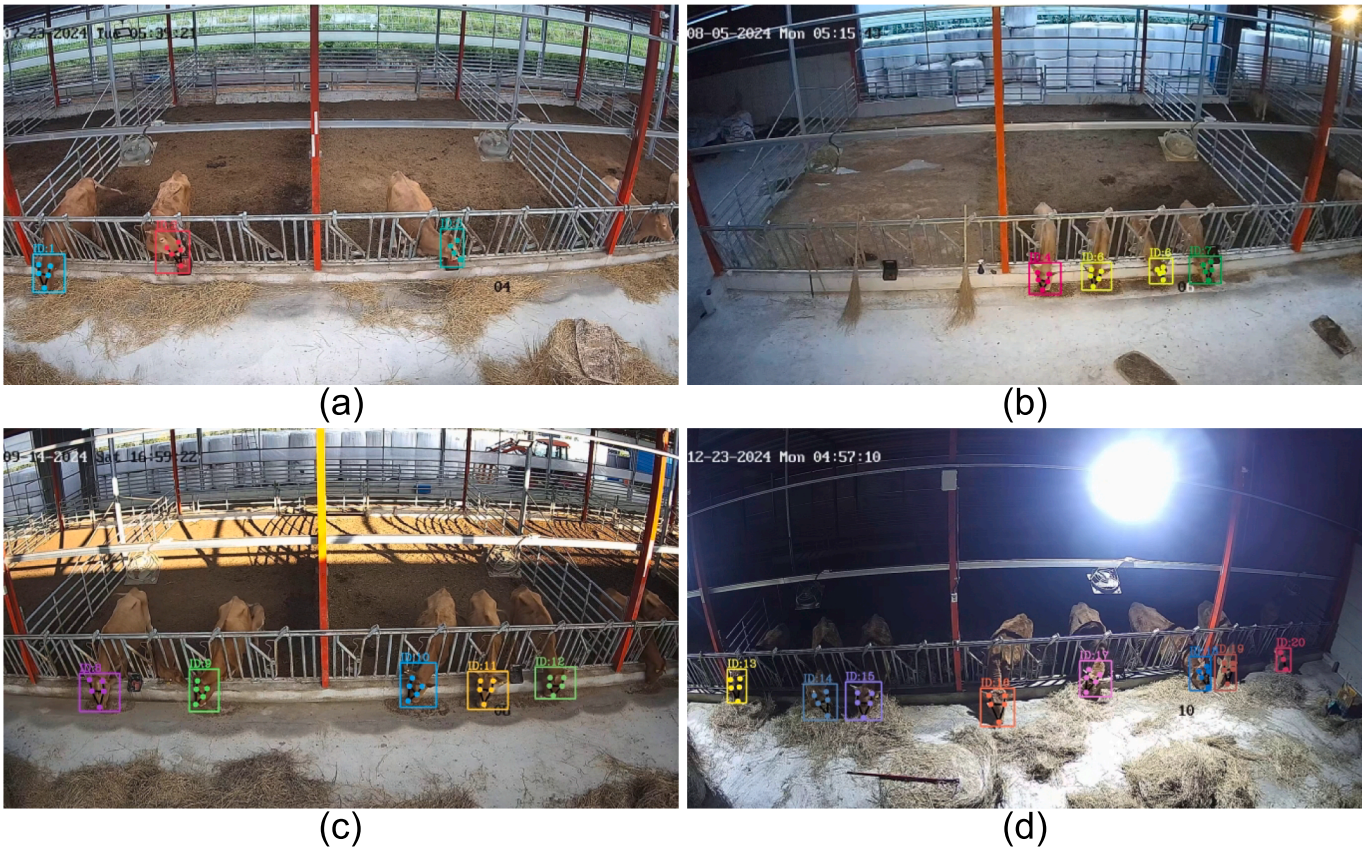


Fig. 17. Quantitative representation of recognition results on the additional dataset. (a) shows the results on July Morning – Adult Individuals data, (b) shows the results on August Morning – Juvenile Individuals data, (c) shows the results on September Afternoon – Adult Individuals data, and (d) shows the results on December Evening – Adult Individuals data.

features a larger number of cattle, comprising fully mature individuals as well as juvenile individuals. Additionally, the environmental variations are more extensive, incorporating both daytime and nighttime scenarios. Our method was fine-tuned only on the additional dataset without extensive training. The ratio of images in the fine-tuning set to those in the validation set is 2:5, maintaining the characteristic of requiring minimal training data.

Table 7 presents the quantitative results of our method on the additional dataset. When evaluated with a relatively lenient metric, such as mAP50, the model demonstrates strong performance. However, as the mAP threshold increases, the model's performance declines significantly. This decline can primarily be attributed to the reduced effectiveness of the detector. As illustrated in the qualitative results in Fig. 17, the model is able to accurately identify the IDs of most cattle, with only one error observed in (b), where two juvenile cattle were misidentified as ID *#6*. However, the bounding boxes exhibit noticeable inaccuracies, particularly for *cattle#2*, *#17*, *#19*, and *#20*.

These observations indicate that, within our method, the detector is more sensitive to data variability compared to the ID classifier, and thus the dependency on data cannot be entirely eliminated. Nevertheless, the process of fine-tuning the model remains relatively straightforward. Our method demonstrates robust adaptability to diverse and complex conditions, such as varying camera perspectives, different times of day, and the introduction of new cattle in a novel farm setting. Particularly in scenarios where high precision is not strictly required, the model maintains a high level of accuracy. These findings underscore the method's capacity to accommodate a wide range of variations in cattle farms, thereby affirming its generalizability and practical applicability.

4. Discussion

Through the combination of farm knowledge, including temporal and pose alignment, illumination augmentation, and semantic segmentation, we achieved promising cattle face recognition performance across extended time spans. Notably, alignment techniques and illumination augmentation yield substantial enhancements to the model's performance. These techniques imposed a consistent domain for data from different dates, greatly enhancing the model's training efficiency. Additionally, the extraction of foreground via the semantic segmentation algorithm not only improved recognition performance but also maintained the model's focus on facial recognition rather than background elements, ensuring its integrity. Collectively, these methods enhanced the model's robustness and generalizability.

Nonetheless, our system still has limitations, as we have not fully utilized all available farm knowledge. For instance, knowing that some cattle cannot appear in two places at the same time could be used to synchronize the IDs of cattle moving across cameras, thereby preventing errors like the one shown in Fig. 13**(b)**. Another example is the dynamic nature of the herd, where some cattle leave and new ones join. In these cases, identifying and storing new IDs, along with establishing an individual information database, is a crucial approach for future research. These methods would allow the cattle population on the farm to expand, enabling the model to recognize more than just 17 cattle. Additionally, they would improve the model's generalizability, ensuring stable performance across different farms and varying camera angles. This would enable the model to adapt to the constantly changing cattle population and accurately reflect real-world working conditions. Advancing our system in this direction is expected to lead to more comprehensive and effective recognition results.

In summary, our framework has proven to be an adaptable solution that is both easy to deploy and update. Its functionality relies solely on computer-based operations, minimizing interference with the routine activities of the cattle.

When deploying the model in practical applications, it is essential to balance computational efficiency and accuracy. If the deployment environment is similar to our farm or constraints are limiting the number of cattle to be identified at a time, it is recommended to prioritize higher accuracy settings. Even if system delays occur due to computational load, tracking cattle can be easily resumed since the number of cattle is relatively small. Conversely, if the number of active cattle on the farm significantly exceeds our setup, prioritizing computational efficiency is advisable to prevent the system from missing detections due to delays. In such cases, skipping modules like FC, FF, and PF can improve computational efficiency. However, skipping IA and TA is not recommended, as these modules have a greater impact on the model's performance.

In the future, observing and utilizing more farm knowledge, continuing to address the limitations encountered in practical applications, collecting more data, expanding the cattle population, and further improving the system's operational efficiency are all meaningful directions for exploration.

5. Conclusions

Recognizing cattle faces in real-world scenarios poses intricate challenges due to the presence of complex constraints such as pose variations, changes in illumination, and background influences. These factors collectively contribute to reducing the generalizability of recognition systems. In this paper, we proposed novel methodologies to address these challenges. Specifically, we advocate for aligning pose through tracking, countering illumination impacts via data augmentation, and isolating the foreground using semantic segmentation. These techniques together helped restrict the sampling environment to real-world conditions, enabling the training of cattle face recognition systems with enhanced generality. Notably, our approach achieved an accuracy of 98.0%, even with a 5-day time gap between the training and test sets. This remarkable consistency underscores the system's robustness and reliability in real-world environments, further demonstrating its efficacy in tackling complex and dynamic scenarios.

Statement on the Use of Generative AI and AI assisted technologies in the writing process

During the preparation of this work the author(s) used "ChatGPT 4o" in order to polish academic writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

CRediT authorship contribution statement

Shujie Han: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alvaro Fuentes:** Writing – review & editing, Investigation, Formal analysis, Conceptualization. **Jongbin Park:** Investigation, Data curation. **Sook Yoon:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Jucheng Yang:** Writing – review & editing, Supervision, Methodology. **Yongchae Jeong:** Supervision. **Dong Sun Park:** Supervision, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00392406); and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(2020R1A2C2013060); and by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2019-NR040079).

Appendix:. COCO evaluation metrics

COCO evaluation metrics rely on Mean Average Precision (mAP) and Mean Average Recall (mAR) to assess model performance in object detection and keypoint detection tasks. The following sections detail the calculation methods and corresponding formulas.

1. Object detection

1.1. Intersection over Union (IoU)

IoU measures the overlap between a predicted bounding box and a ground truth box. It is defined as:

$$IoU = \frac{A_{overlap}}{A_{union}}$$

where $A_{overlap}$ is the area of the intersection between the predicted box and the ground truth box; A_{union} is the area of their union. A prediction is considered correct if the IoU exceeds a specified threshold (e.g., 0.5 or 0.75).

*1.2. **Mean Average precision** (mAP)*

1.2.1. Per-Class Average precision

For a specific class c, the Average Precision (AP) at a fixed IoU threshold is calculated as:

$$AP_{IoU}^{(c)} = \frac{1}{|\mathscr{R}|}\sum_{r\in\mathscr{R}} P_c(r)$$

where $P_c(r)$ is the precision for class c at recall r; $\mathscr{R}$ is the set of sampled recall values from 0 to 1.

1.2.2. Overall mAP

The mAP at a specific IoU threshold is obtained by averaging the AP over all classes:

$$mAP_{IoU} = \frac{1}{N_c}\sum_{c=1}^{N_c} AP_{IoU}^{(c)}$$

where N_c is the total number of classes.

1.3. Evaluation metrics used in this study

- mAP50: mAP computed at IoU = 0.5.
- mAP75: mAP computed at IoU = 0.75.
- mAP50:95: mAP averaged over IoU thresholds from 0.5 to 0.95 (in steps of 0.05).

$$mAP50:95 = \frac{1}{10}\sum_{IoU=0.5}^{0.95} mAP_{IoU}$$

1.4. Mean Average recall (mAR)

The mAR measures the model's ability to detect ground truth instances under different recall conditions. For a specific class c, the Average Recall (AR) is calculated as:

$$AR(c) = \frac{1}{|\mathscr{T}|}\sum_{IoU\in\mathscr{T}} R_c(IoU)$$

where $\mathscr{T}$ is the set of IoU thresholds (e.g., $\{0.5, 0.55, \cdots, 0.95\}$); $R_c(IoU)$ is the recall for class c at a given IoU. The overall mAR is then averaged across all classes:

$$mAR = \frac{1}{N_c}\sum_{c=1}^{N_c} AR(c)$$

2. Keypoint detection

2.1. Object keypoint similarity (OKS)

The OKS measures the similarity between predicted keypoints and ground truth keypoints, defined as:

$$OKS = \frac{\sum_i \left[exp\left(-\frac{d_i^2}{2s^2k_i^2} \right) \cdot \delta(v_i > 0) \right]}{\sum_i \delta(v_i > 0)}$$

where d_i Euclidean distance between the predicted and ground truth keypoints; s area of the bounding box of the object; k_i normalization factor for each keypoint; v_i visibility flag of the keypoint ($v_i > 0$ indicates visibility); $\delta(v_i > 0)$: indicator function that evaluates to 1 if $v_i > 0$ and 0 otherwise.

2.2. Mean Average precision (mAP)

2.2.1. Per-Class Average precision

For a specific class c, the AP at a fixed OKS threshold is:

$$AP_{OKS}^{(c)} = \frac{1}{|\mathscr{R}|}\sum_{r\in\mathscr{R}} P_c(r)$$

2.2.2. Overall mAP

The mAP at a specific OKS threshold is calculated as:

$$mAP_{OKS} = \frac{1}{N_c}\sum_{c=1}^{N_c} AP_{OKS}^{(c)}$$

2.3. Evaluation metrics used in this study

- mAP50: mAP computed at OKS = 0.5.
- mAP75: mAP computed at OKS = 0.75.
- mAP50:95: mAP averaged over OKS thresholds from 0.5 to 0.95 (in steps of 0.05).

$$mAP50:95 = \frac{1}{10}\sum_{OKS=0.5}^{0.95} mAP_{OKS}$$

3. Precision and recall

Precision quantifies the proportion of correctly predicted instances among all predicted instances:

$$P = \frac{TP}{TP + FP}$$

where TP is true positives; FP is false positives.

Recall quantifies the proportion of correctly predicted instances among all ground truth instances:

$$R = \frac{TP}{TP + FN}$$

where FN is false negatives.

Data availability

Data will be made available on request.

References

An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T., 2022. Killing Two Birds with One Stone: Efficient and Robust Training of Face Recognition CNNs by Partial FC, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4032-4041.

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016. Simple online and realtime tracking, 2016 IEEE international conference on image processing (ICIP). IEEE 3464–3468.

Cao, J., Tang, H., Fang, H., Shen, X., Tai, Y., Lu, C., 2019. Cross-Domain Adaptation for Animal Pose Estimation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society, Los Alamitos, CA, USA, pp. 9497–9506.

Chen, X., Yang, T., Mai, K., Liu, C., Xiong, J., Kuang, Y., Gao, Y., 2022. Holstein cattle face re-identification unifying global and part feature deep network with attention mechanism. Animals 12, 1047.

de Melo, B.A., Nascimento, I.D.M., Santos, L.T.A.D., de Lima, L.G., de Araújo, F.C.T., Rios, R.R.S., Couto, A.D.G., Fraga, A.B., 2018. Body morphometric measurements in Murrah crossbred buffaloes (Bubalus bubalis). J. Appl. Anim. Res. 46 (1), 1307–1312.

Du, Z., Shi, M., Deng, J., 2023. Boosting object detection with zero-shot day-night domain adaptation. arXiv preprint arXiv:2312.01220.

Fuentes, A., Han, S., Nasir, M.F., Park, J., Yoon, S., Park, D.S., 2023. Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning. Animals 13, 2020.

Han, S., Fuentes, A., Yoon, S., Jeong, Y., Kim, H., Park, D.S., 2023. Deep learning-based multi-cattle tracking in crowded livestock farming using video. Comput. Electron. Agric. 212, 108044.

Han, S., Fuentes, A., Yoon, S., Jeong, Y., Park, D.S., 2022. Cattle Face Pose Estimation using Landmark Points for Smart Livestock Farming, International Conference on Smartmedia (SMA), Saipan, USA.

Hao, W., Zhang, K., Han, M., Hao, W., Wang, J., Li, F., Liu, Z., 2023. A novel Jinnan individual cattle recognition approach based on mutual attention learning scheme. Expert Syst. Appl. 120551.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770-778.

Hewitt, C., Mahmoud, M., 2019. Pose-Informed Face Alignment for Extreme Head Pose Variations in Animals, 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1-6.

Hu, H., Dai, B., Shen, W., Wei, X., Sun, J., Li, R., Zhang, Y., 2020. Cow identification based on fusion of deep parts features. Biosyst. Eng. 192, 245–256.

Jaddoa, M., Gonzalez, L., Cuthbertson, H., Al-Jumaily, A., 2019. Multi View Face Detection in Cattle Using Infrared Thermography. In: International Conference on Applied Computing to Support Industry: Innovation and Technology. Springer, pp. 223–236.

Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G., 2020. Animalweb: A large-scale hierarchical dataset of annotated animal faces, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6939-6948.

Larregui, J.I., Cazzato, D., Castro, S.M., 2019. An image processing pipeline to segment iris for unconstrained cow identification system. Open Comput. Sci. 9, 145–159.

Li, G., Erickson, G.E., Xiong, Y., 2022a. Individual beef cattle identification using muzzle images and deep learning techniques. Animals 12, 1453.

Li, N., Ren, Z., Li, D., Zeng, L., 2020. Automated techniques for monitoring the behaviour and welfare of broilers and laying hens: towards the goal of precision livestock farming. animal 14 (3), 617–625.

Li, Y., Yang, S., Liu, P., Zhang, S., Wang, Y., Wang, Z., Yang, W., Xia, S.-T., 2022b. SimCC: A Simple Coordinate Classification Perspective for Human Pose Estimation. In: Part, V.I. (Ed.), Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings. Springer, pp. 89–106.

Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L., 2022. Dab-detr: dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329.

Meng, Q., Gao, X., Wen, Y., Fang, X., Su, X., 2024. Few-shot Classification with Shared and Private Branches for Cow Face Recognition. In: IEEE, pp. 1–5.

Meng, Y., Yoon, S., Han, S., Fuentes, A., Park, J., Jeong, Y., Park, D.S., 2023. Improving known–unknown cattle's face recognition for smart livestock farm management. Animals 13.

Neethirajan, S., 2021. Happy cow or thinking pig? Wur wolf-facial coding platform for measuring emotions in farm animals. AI 2, 342–354.

Pretto, A., Savio, G., Gottardo, F., Uccheddu, F., Concheri, G., 2022. A novel low-cost visual ear tag based identification system for precision beef cattle livestock farming. Information Process. Agric.

Qiao, L., Geng, Y., Zhang, Y., Zhang, S., Xu, C., 2022. MobiCFNet: A Lightweight Model for Cattle Face Recognition in Nature, Intelligence Science IV: 5th IFIP TC 12 International Conference, ICIS 2022, Xi'an, China, October 28–31, 2022, Proceedings. Springer, pp. 386-394.

Qiao, Y., Guo, Y., He, D., 2023. Cattle body detection based on YOLOv5-ASFF for precision livestock farming. Comput. Electron. Agric. 204, 107579.

Qiao, Y., Kong, H., Clark, C., Lomax, S., Su, D., Eiffert, S., Sukkarieh, S., 2021. Intelligent perception for cattle monitoring: a review for cattle identification, body condition score evaluation, and weight estimation. Comput. Electron. Agric. 185, 106143.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28.

Ruchay, A., Akulshin, I., Kolpakov, V., Dzhulamanov, K., Guo, H., Pezzuolo, A., 2023. Cattle Face Recognition Using Deep Transfer Learning Techniques. In: IEEE, pp. 569–574.

Ruchay, A., Kolpakov, V., Guo, H., Pezzuolo, A., 2024. On-barn cattle facial recognition using deep transfer learning and data augmentation. Comput. Electron. Agric. 225, 109306.

Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y., 2018. Integral human pose regression, Proceedings of the European conference on computer vision (ECCV), pp. 529-545.

Ter-Sarkisov, A., Ross, R., Kelleher, J., 2017. Bootstrapping labelled dataset construction for cow tracking and behavior analysis. In: 2017 14th Conference on Computer and Robot Vision (CRV). IEEE, pp. 277–284.

Wang, C., Bochkovskiy, A., Liao, H., 2023. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In: 2023 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pp. 7464–7475.
Wang, M., Deng, W., 2021. Deep face recognition: a survey. Neurocomputing 429, 215–244.
Wang, Z., Ni, F., Yao, N., 2020. Mtfcn: multi-task fully convolutional network for cow face detection. In: Springer, pp. 1116–1127.
Wei, S.-E., Ramakrishna, V., Kanade, T., Sheikh, Y., 2016. Convolutional pose machines, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4724-4732.
Wu, X., Kawanishi, T., Kashino, K., 2020. Reflectance-Guided, Contrast-Accumulated Histogram Equalization. In: IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 2498–2502.
Xu, B., Wang, W., Guo, L., Chen, G., Wang, Y., Zhang, W., Li, Y., 2021. Evaluation of deep learning for automatic multi-view face detection in cattle. Agriculture 11, 1062.
Xu, X., Deng, H., Wang, Y., Zhang, S., Song, H., 2024. Boosting cattle face recognition under uncontrolled scenes by embedding enhancement and optimization. Appl. Soft Comput. 164, 111951.
Yao, L., Hu, Z., Liu, C., Liu, H., Kuang, Y., Gao, Y., 2019. Cow face detection and recognition based on automatic feature extraction algorithm, Proceedings of the ACM turing celebration conference-china, pp. 1-5.
Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J., 2021. Lite-hrnet: A lightweight high-resolution network, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10440-10450.
Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y., 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.
Zhang, X., Wan, F., Liu, C., Ji, R., Ye, Q., 2019. Freeanchor: Learning to match anchors for visual object detection. Advances in neural information processing systems 32.
Zhang, Z., Gao, J., Xu, F., Chen, J., 2023. Siamese GC capsule networks for small sample cow face recognition. IEEE Access.