agriculture

MDPI

*Article*

# Knowledge-Enhanced Deep Learning for Identity-Preserved Multi-Camera Cattle Tracking

**Shujie Han [1,2,†], Alvaro Fuentes [1,2,†], Jiaqi Liu [1,2], Zihan Du [3], Jongbin Park [1,2], Jucheng Yang [4], Yongchae Jeong [5], Sook Yoon [6,*] and Dong Sun Park [1,2,*]**

[1] Department of Electronics Engineering, Jeonbuk National University, Jeonju 54896, Republic of Korea; shujiejulie@jbnu.ac.kr (S.H.); afuentes@jbnu.ac.kr (A.F.); ljq9655@gmail.com (J.L.); jbpark99@jbnu.ac.kr (J.P.)
[2] Core Research Institute of Intelligent Robots, Jeonbuk National University, Jeonju 54896, Republic of Korea
[3] College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China; 13483211514@163.com
[4] School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541000, China; jcyang@guet.edu.cn
[5] Division of Electronics and Information Engineering, IT Convergence Research Center, Jeonbuk National University, Jeonju 54896, Republic of Korea; ycjeong@jbnu.ac.kr
[6] Department of Computer Engineering, Mokpo National University, Muan 58554, Republic of Korea
* Correspondence: syoon@mokpo.ac.kr (S.Y.); dspark@jbnu.ac.kr (D.S.P.)
† These authors contributed equally to this work.

## Abstract

Accurate long-term tracking of individual cattle is essential for precision livestock farming but remains challenging due to occlusions, posture variability, and identity drift in free-range environments. We propose a multi-camera tracking framework that combines bird's-eye-view (BEV) trajectory matching with cattle face recognition to ensure identity preservation across long video sequences. A large-scale dataset was collected from five synchronized 4K cameras in a commercial barn, capturing both full-body movements and frontal facial views. The system employs center point detection and BEV projection for cross-view trajectory association, while periodic face recognition during feeding refreshes identity assignments and corrects errors. Evaluations on a two-day dataset of more than 600,000 images demonstrate robust performance, with an AssPr of 84.481% and a LocA score of 78.836%. The framework outperforms baseline trajectory matching methods, maintaining identity consistency under dense crowding and noisy labels. These results demonstrate a practical and scalable solution for automated cattle monitoring, advancing data-driven livestock management and welfare.

**Keywords:** multi-camera tracking; bird's-eye-view projection; face recognition; cross-view identity association; large-scale imperfect datasets

check for updates

Academic Editors: Yang Zhao and Beibei Xu

Received: 23 July 2025
Revised: 4 September 2025
Accepted: 16 September 2025
Published: 18 September 2025

**Citation:** Han, S.; Fuentes, A.; Liu, J.; Du, Z.; Park, J.; Yang, J.; Jeong, Y.; Yoon, S.; Park, D.S. Knowledge-Enhanced Deep Learning for Identity-Preserved Multi-Camera Cattle Tracking. *Agriculture* **2025**, *15*, 1970. https://doi.org/10.3390/agriculture15181970

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Accurate recognition and behavioral tracking of individual cattle are crucial for timely disease detection and prevention [1], improved production efficiency, and enhanced consumer safety [2]. Traditional identification methods, such as wearable devices, ear tags, and RFID systems, have notable drawbacks, including interference with natural behavior, high implementation costs, and susceptibility to physical damage [3–6]. In contrast, computer vision technologies offer promising alternatives due to their low operational costs, ease of deployment, and compatibility with animal welfare principles [7].

Conventional monitoring systems typically rely on single-camera setups [8], which are restricted to small environments where cattle movement is constrained. With the growing adoption of animal welfare standards and the transition toward free-range farming, however, cattle now exhibit greater freedom of movement, introducing substantial postural variability [9]. Despite advancements in object detection and tracking, existing approaches still face challenges in real-world barns due to occlusions and variations in body posture [10]. Combined with feature drift, these factors significantly degrade the accuracy and robustness of single-camera systems.

To overcome these limitations, multi-camera tracking technologies have gained increasing attention for tasks involving individual cattle recognition and long-term behavior monitoring. By leveraging multi-view data fusion, such systems effectively address issues related to occlusion and posture variation while reducing blind spots in coverage [11]. Standard multi-object tracking pipelines typically involve two stages: generating local trajectories within each camera's field of view and then integrating these trajectories into coherent global paths through cross-camera association [12,13].

Recent research has explored both end-to-end frameworks and trajectory matching approaches for multi-camera tracking [14–17]. These methods address cross-camera trajectory association through strategies such as bird's-eye-view (BEV) alignment, constraint-augmented trajectory matching, and graph-based integration of appearance and motion cues. However, applying these techniques to cattle remains particularly challenging. For example, Zheng et al. [18] proposed a YOLOv7–ByteTrack-based approach that excluded re-identification modules to reduce costs, but it still suffered from frequent identity switches in complex environments [19,20]. Guo et al. [21] improved livestock re-identification performance through fusion strategies [22,23], yet frequent ID switches persisted [24]. Nguyen et al. [8] noted that multi-stage pipelines may propagate local identity errors to global results, while Myat et al. [2] attempted to track contour-based features [25] but encountered trajectory fragmentation in long-term tracking. Chen et al. [7] further demonstrated that cattle face recognition using re-identification techniques enables cross-scenario generalization by learning discriminative features, offering an effective solution for identity verification in new settings.

Previous studies confirm that relying solely on appearance features is risky, given cattle's large body size and posture variability; positional cues alone also lead to errors from missed detections, and long-term tracking often suffers from trajectory interruptions [26–28]. In our experimental barn, five cameras were deployed at different viewpoints (Figure 1), revealing three key challenges: (i) occlusion among individuals, particularly during feeding; (ii) identity switches caused by posture changes both within and across views; and (iii) appearance drift between action-view surveillance and face-view feeding perspectives, which complicates cross-view association.

To overcome these issues, we propose a multi-faceted framework for cross-camera cattle tracking. First, we annotate each animal's center of mass (center points) and train a model to detect these ground-projected keypoints. Representing trajectories as center point sequences reduces errors from relying solely on position or appearance. Second, detected center points are projected onto a BEV map, where a location-based trajectory matching algorithm ensures consistent association; each camera is mapped onto a distinct, non-overlapping BEV region to prevent ID switches. Finally, during feeding, we refresh and correct IDs using cattle face recognition, enabling robust long-term tracking. This multi-pronged strategy exploits spatial information and cross-view consistency, circumventing the limitations of appearance-based methods (Figure 2).

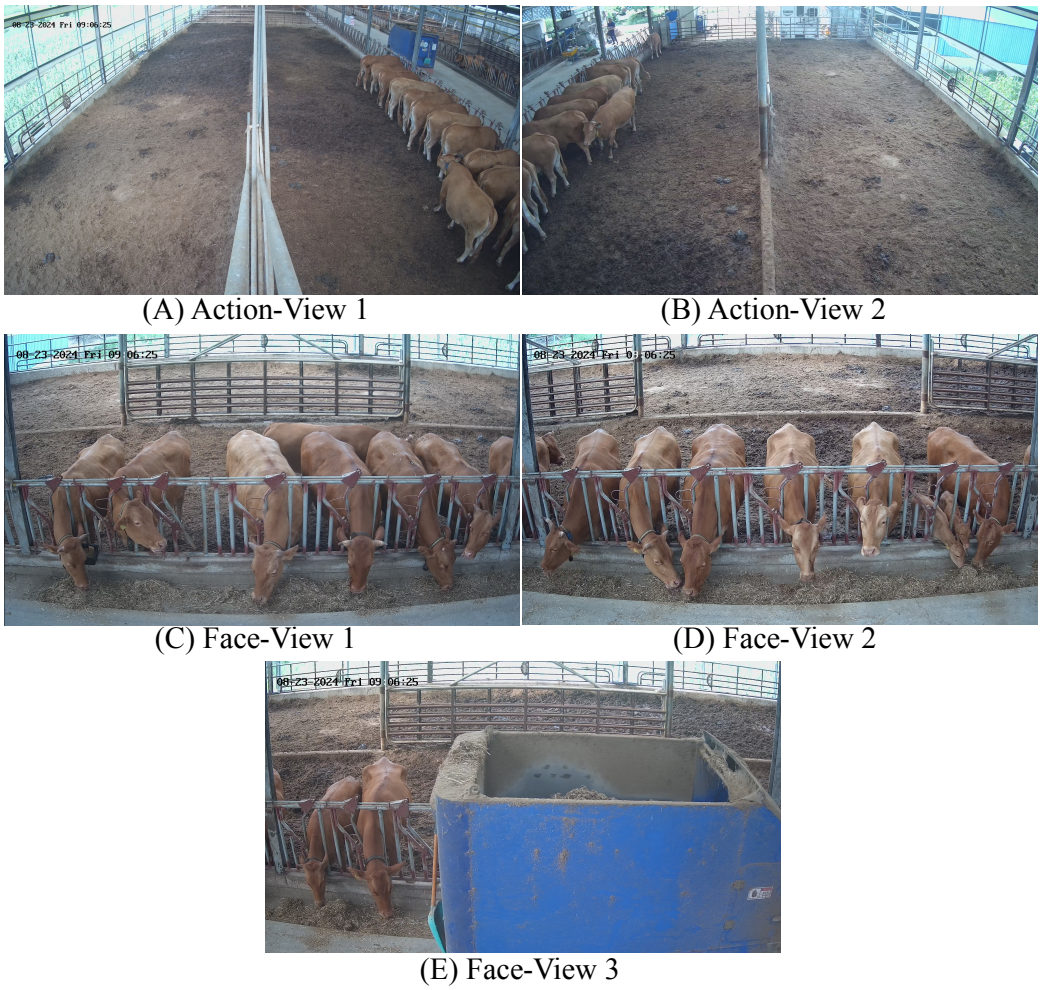



**Figure 1.** Multi-view surveillance setup in the cattle feeding area. (**A**,**B**) Action views captured from overhead cameras to observe cattle movement and behavioral patterns. (**C**–**E**) Face views captured from frontal cameras to enable face recognition.

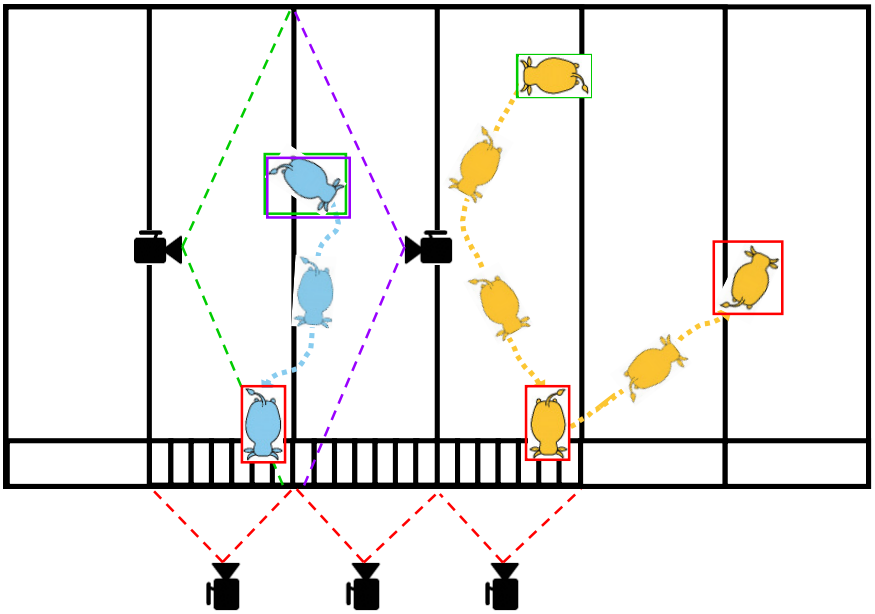


**Figure 2.** Schematic illustration of the multi-view trajectory association framework. The figure demonstrates the trajectory matching strategies across action–action cameras and across action–face cameras on a unified BEV map.

Specifically, we collected a two-day dataset with continuous surveillance from five cameras: two action-view cameras for movement and three face-view cameras for identity during feeding. Models were trained to jointly detect bodies, center points, faces, and landmarks under multi-view conditions. Tracking proceeded in two stages: cross-action-view trajectory matching and cross-face-view ID assignment. For action views, center points were projected onto the BEV and matched spatially. For face views, we employed the Knowledge-Based Cattle Face Recognition (KBCFR) model [29], which outputs up to five candidate IDs per face. These identities are recursively propagated across cameras, and BEV projections are divided into a $16 \times 42$ grid, providing spatial cues for linking face-view and action-view trajectories.

Our dataset is substantial, comprising 615,963 images captured from multiple viewpoints. To improve annotation efficiency, we adopted a noise-labeling strategy: in action views, precise annotations were provided only during action transitions, while labels for other frames were auto-generated from these reference points. In face views, IDs were first predicted using the pretrained KBCFR model and then manually corrected for errors and ambiguities. Despite the presence of noisy labels, our results demonstrate that models trained on this large-scale dataset achieve strong performance. Specifically, training on the first day and evaluating on the second day yielded an average AssPr of 84.481% across multiple views, together with a LocA score of 78.836%. These findings confirm the effectiveness of our method for accurate and persistent cattle tracking under real-world farm conditions.

## 2. Materials and Methods

### *2.1. Data Acquisition*

To collect authentic livestock farming data, and following deployment strategies proposed in previous studies [30,31], we installed five Hikvision 4K ultra-high-definition surveillance cameras (Hikvision Digital Technology Co., Ltd., Hangzhou, China) ($3840 \times 2160$ resolution) in a closed Hanwoo cattle barn (12 m × 36 m) located in Imsil, South Korea. The barn was equipped with anti-slip concrete flooring layered with sawdust bedding, semi-transparent polycarbonate roofing, and an active ventilation system. It housed 17 Hanwoo cattle aged 1–7 years, excluding 2 calves.

The camera system comprised two subsystems: (i) two action-view cameras mounted at opposite ends of the barn roof at a height of 8.5 m with a 30° tilt angle covering the entire barn for motion tracking; and (ii) three face-view cameras installed 6.6 m opposite the feed barrier, each monitoring seven contiguous feeding stalls (0.83 m × 0.62 m per space), jointly covering a total of 21 stalls. Multi-view video data were continuously recorded on 23–24 August 2024. The original 15 fps recordings were downsampled to 1 fps and resized to $1920 \times 1080$ resolution.

Table 1 presents a breakdown of the collected multi-view dataset. A total of 367,494 frames were captured across the two days, covering both daytime and nighttime conditions. The dataset includes images from all five camera views: two action-view and three face-view cameras.

The training set consisted of data collected on 23 August 2024, supplemented with additional face-view samples recorded in 2021 under identical camera configurations. In total, the training subset included 164,027 action-view frames and 29,936 face-view frames. The supplementary 2021 data added 24,194 face-view images (8194, 9000, and 7000 frames from Face-View 1, 2, and 3, respectively), increasing the diversity and balance of facial samples for recognition tasks.

The testing set comprised 168,800 action-view frames and 4731 face-view frames collected on 24 August 2024. These samples were used to evaluate tracking performance

under unconstrained conditions, where face recognition was performed whenever an animal appeared in the face cameras, regardless of feeding behavior.

**Table 1.** Details of the multi-view dataset for cattle tracking. Each number represents the number of frames extracted from the video.

| Date and Time | Action-View 1 | Action-View 2 | Face-View 1 | Face-View 2 | Face-View 3 |
|---|---|---|---|---|---|
| 23 August 2024 (Daytime) | 53,813 | 49,777 | 523 | 523 | 523 |
| 23 August 2024 (Nighttime) | 28,207 | 32,330 | 1391 | 1391 | 1391 |
| Additional (2021, Daytime) | – | – | 8194 | 9000 | 7000 |
| **Training Total** | **164,027** | | | **29,936** | |
| 24 August 2024 (Daytime) | 49,432 | 47,999 | 1089 | 1089 | 1089 |
| 24 August 2024 (Nighttime) | 34,968 | 36,401 | 488 | 488 | 488 |
| **Testing Total** | **168,800** | | | **4731** | |

The dataset annotations include bounding boxes, center points, IDs, and facial keypoints. Bounding boxes mark the positions of cattle bodies or faces, with IDs assigned simultaneously. Importantly, IDs are unified across action and face views: the same animal is consistently given the same ID regardless of camera perspective. While the five facial keypoints are relatively straightforward to annotate, the center point requires more manual estimation. Following our previous work [28], the center point is defined as the ground projection of the body's midpoint located between the four limbs and aligned with the spinal vertebra.

To meet the demands of multi-view annotation, we adopted a hybrid strategy. In action views, manual annotation was performed only during transitions between actions (e.g., walking to standing), while labels for periods of consistent actions were automatically propagated as noise labels from the previously annotated frames. In face views, a pretrained KBCFR model was used to generate initial bounding boxes and facial keypoints (nose, eyes, and ear bases). These annotations were then verified and manually refined to correct labeling errors and ID mismatches. Examples of annotation results are shown in Figure 3, where most noise labels remain reasonably accurate. Annotations with clearly visible errors were directly removed from the training data to prevent negative impacts on model performance.




(A) Action-View annotation example (B) Face-View annotation example

**Figure 3.** Annotation examples in different camera views. (**A**) Action-view annotation example with bounding boxes and center points labeled for body tracking. (**B**) Face-view annotation example with facial bounding boxes and landmarks labeled for individual identification.

### *2.2. Method*

In this section, we present the framework developed for robust multi-view cattle tracking. We begin with an overview of the system architecture, highlighting its design principles

and key components. Next, we describe the Action-View Trajectory Matching method, which aligns trajectories based on spatial positions in the BEV map. We then introduce the Face-View ID Assignment approach, which employs the KBCFR model to improve identity recognition. Finally, we detail the Cross-View ID Update mechanism, which integrates results from multiple views to refine and unify identity tracking across modalities.

#### 2.2.1. Overview

Cross-view cattle tracking is essential, as broader coverage and occlusion handling require observations from multiple complementary viewpoints. Figure 4 illustrates the proposed framework, highlighting its main components and workflow. During inference, the framework processes synchronized multi-view video frames from five cameras. To represent both spatial and temporal information, we construct a 5D input tensor of shape `(batch, views, channels, width, height)`. The `batch` dimension encodes temporal continuity, while the `views` dimension indexes the five perspectives. This setup enables tracking across time and facilitates cross-view association.

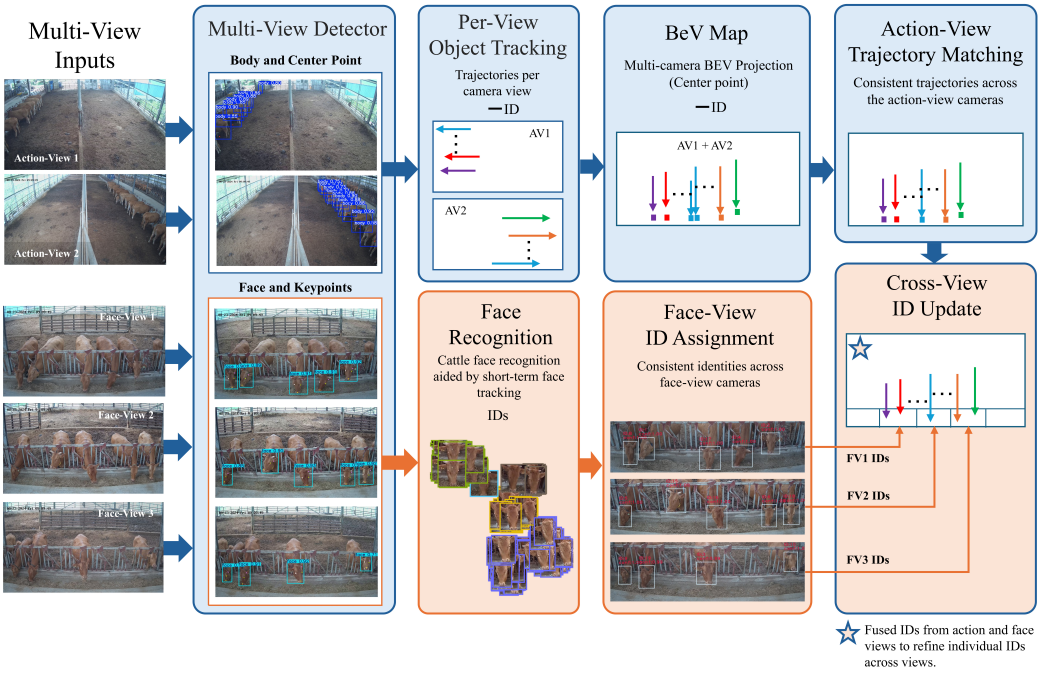



**Figure 4.** The workflow of the proposed cattle tracking system. The system first takes multi-view images as Inputs, which are processed by the detector to identify cattle in both action and face views. The Tracker generates temporal trajectories from action views, which are further projected onto a BEV map for spatial alignment. The Action-View Trajectory Matching module performs trajectory association across views. In parallel, the Face Recognition module extracts facial features and recognizes identities, then Face-View ID Assignment labels individual cattle. Finally, the Cross-View ID Update integrates trajectory and face information to maintain consistent identities across all views.

YOLOv11x-pose [32] is first applied to detect cattle bodies and faces. The detector outputs bounding boxes and keypoints but does not assign identities. In action views, the keypoint corresponds to the cattle's center point projected onto the ground plane, while in face views, five landmarks are detected (nose, eyes, and ear bases). These outputs provide candidate features for tracking, localization, alignment, and identification.

Detection results are processed independently per camera. In action views, ByteTrack (Ultralytics v8.3.3) is used to generate local trajectories. These trajectories are local and

view-specific, meaning the same animal may appear in multiple views with different temporary IDs (illustrated as blue paths in Figure 4).

In face views, the KBCFR model identifies individuals by tracking faces across short sequences and selecting representative samples for recognition. To address uncertainty, the model returns up to five ranked candidate IDs per trajectory to support ID correction, which is especially useful during feeding when body tracking is unreliable but facial views remain accurate.

To enable cross-view association, we construct a global BEV map of the barn. Action-view trajectories are projected onto this map. Given the visual similarity among cattle, spatial rather than appearance features are used for trajectory matching. The Action-View Trajectory Matching module estimates similarity by comparing occupancy in BEV grids.

For Face-View ID Assignment, we enforce the constraint that the same animal cannot appear in multiple stalls simultaneously. Ranked predictions from the KBCFR model guide conflict resolution, ensuring unique global IDs across face views. Finally, the Cross-View ID Update module aligns face IDs with action-view trajectories in the feeding zone, where cattle appear most reliably. Only individuals with stable trajectories in this region are eligible for identity fusion, ensuring consistent long-term tracking.

Details of each component of the system are presented in the following sections.

#### 2.2.2. Action-View Trajectory Matching

The Action-View Trajectory Matching module aligns local trajectories from multiple action-view cameras into global trajectories on the BEV map. The process begins by projecting detected cattle center points from each camera onto the BEV using homography transformation. Each trajectory is represented as a sequence of center points, enabling alignment across cameras based on spatial overlap in the BEV grid. The BEV region is defined as a rectangular area measuring 12 m $\times$ 36 m. To determine trajectory locations on the BEV map, the area is divided into 14 $\times$ 42 grids. The grid size is based on the feed fence dimensions: in the real world, seven feeding positions span every 6 m, so the size of one feeding stall is used as the grid unit. An example of the resulting BEV map is shown in Figure 5.

The BEV projection is performed using a homography transformation, which leverages the planar geometry between the image plane and the ground plane. For each calibrated camera, a homography matrix $\mathbf{H} \in \mathbb{R}^{3\times3}$ is defined to map coordinates from the ground plane to the image plane. To reverse-transform image coordinates back to ground coordinates, we compute the inverse matrix $\mathbf{H}^{-1}$ and apply it to image pixel coordinates.

Let the image pixel coordinate be :

$$\mathbf{p}_{\text{img}} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

Then the corresponding ground-plane coordinate is computed as:

$$\mathbf{p}_{\text{gnd}} = \mathbf{H}^{-1} \cdot \mathbf{p}_{\text{img}} \tag{2}$$

The actual ground-space coordinates are extracted as:

$$X = \mathbf{p}_{\text{gnd}}[0], \quad Y = \mathbf{p}_{\text{gnd}}[1] \tag{3}$$

Toconvert ground-plane coordinates to BEV pixel coordinates, we introduce a resolution scale $s$ (pixels per meter) that is set to $s = 100$ pixels/m. Given the ground height $H_m$, the BEV pixel location is computed as:

$$b_x = \lfloor X \cdot s \rfloor, \quad b_y = \lfloor (H_m - Y) \cdot s \rfloor \tag{4}$$

This ensures that the BEV origin is aligned with the bottom-left corner of the physical ground plane, with the $Y$ axis flipped to match image coordinate conventions. The homography matrices for Action-View 1 and 2, denoted $\mathbf{H}_{A1}$ and $\mathbf{H}_{A2}$, respectively, are obtained via offline calibration:

$$\mathbf{H}_{A1} = \begin{bmatrix} -354.892865 & 501.606593 & -1970.1837 \\ 56.4533934 & 13.3831183 & -3014.58171 \\ -0.373157365 & -0.00131341768 & 1.0 \end{bmatrix} \tag{5}$$

$$\mathbf{H}_{A2} = \begin{bmatrix} -40.3048191 & 66.4094257 & 493.0 \\ 10.2632195 & -2.38005941 & 162.0 \\ -0.0462612280 & 0.000908640865 & 1.0 \end{bmatrix} \tag{6}$$

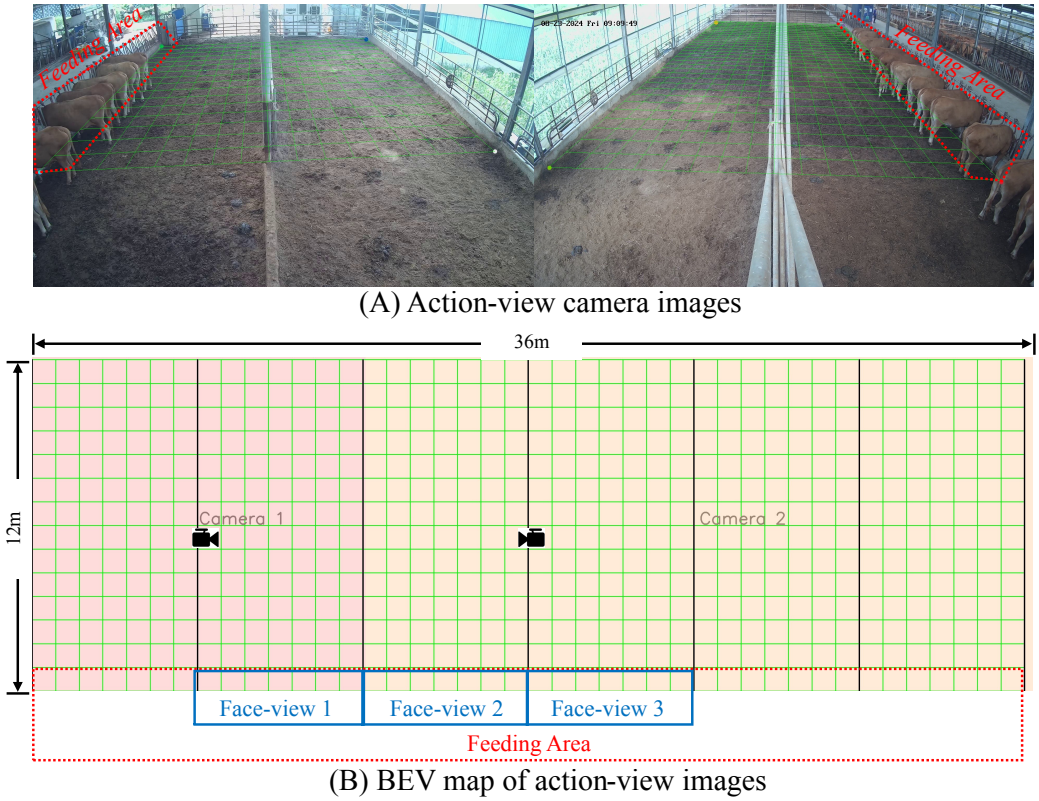



**Figure 5.** Grid-based spatial partitioning and BEV projection overlay. The upper part shows original views from Action-View 1 (Camera 1) and Action-View 2 (Camera 2) overlaid with a uniform grid and labeled Feeding Area. The bottom part displays the corresponding BEV grid projection, aligning both camera views within a unified spatial layout.

Trajectory IDs are robustly matched across views based on grids within the BEV map. Once all observed trajectories are projected onto the BEV coordinate system, we discretize their positions by assigning each trajectory to a specific grid based on its location. Given a predefined BEV partition of $14 \times 42$ uniformly sized grids, each with dimensions $s \times s$ where $s = 85$, $G_i$ denotes the grid to which the endpoint of a trajectory belongs. It is determined from the BEV coordinates $(b_x, b_y)$ according to the following:

$$G_i = \lambda \left\lfloor \frac{b_y}{s} \right\rfloor + \left\lfloor \frac{b_x}{s} \right\rfloor \tag{7}$$

where $\lambda = 42$ presents the number of grids per row on the BEV map.

Let $ID_{G_i}^{A1}$ and $ID_{G_j}^{A2}$ be the trajectory IDs from Action-View 1 and 2 projected onto grids $G_i$ and $G_j$, respectively. The global trajectory ID, denoted as $\widetilde{ID}_G$, is updated based on the BEV positions of $ID_{G_i}^{A1}$ and $ID_{G_j}^{A2}$. To this end, we first define a cumulative global ID counter $C^{(\tau)}$. Then, by directly using $ID_{G_i}^{A1}$ as the global ID, we obtain:

$$\widetilde{ID}_{G_i} = ID_{G_i}^{A1}, \tag{8}$$

with the cumulative counter initialized as:

$$C^{(0)} = |\{ID_G^{A1}\}|, \tag{9}$$

where $\{ID_G^{A1}\}$ denotes the set of trajectory IDs in Action-View 1. Next, the updates of $ID_{G_j}^{A2}$ and $C^{(\tau)}$ are given by:

$$\widetilde{ID}_{G_j} = \begin{cases} ID_{G_i}^{A1}, & G_j = G_i, \\ ID_{G_j}^{A2} + C^{(\tau-1)}, & G_j \neq G_i, \end{cases} \tag{10}$$

$$C^{(\tau)} = \begin{cases} C^{(\tau-1)}, & G_j = G_i, \\ C^{(\tau-1)} + 1, & G_j \neq G_i. \end{cases} \tag{11}$$

This grid-based assignment exploits the physical design of the barn: each stall corresponds to one grid, so at any time, a grid should contain at most one animal. By enforcing this constraint, trajectories that fall into the same grid are highly likely to belong to the same individual, enabling reliable identity association across views.

#### 2.2.3. Face-View ID Assignment

The Face-View ID Assignment strategy improves identity recognition by exploiting frontal facial features captured during feeding. Recognition is performed using our previously developed KBCFR model, which enables stable short-term face tracking and identification through a combined data- and knowledge-driven design.

Cattle face recognition is inherently challenging due to pose variation, illumination changes, and background clutter. To address these challenges, KBCFR integrates three key strategies: (i) selecting the most representative facial instance within a tracked sequence using prior knowledge, (ii) applying illumination-based data augmentation to improve robustness, and (iii) employing instance segmentation to focus on the animal's face while suppressing background interference.

A limitation of KBCFR is its independent treatment of each trajectory, which can cause identity conflicts, such as two different cattle being assigned the same ID. This issue is particularly severe in multi-camera scenarios, where the model lacks cross-view priors and cannot exploit confirmed identities from other camera viewpoints, increasing the likelihood of ID conflicts.

To overcome this problem, we introduce a Face-View ID Assignment mechanism as shown in Algorithm 1 . For each trajectory, the KBCFR model outputs a rank-5 prediction, returning up to five candidate IDs ordered by confidence. Based on the barn's fixed spatial layout, recognition results from the three cameras are concatenated in a predefined order: Face-View 1 followed by 2 and 3. The concatenated ID queue is then checked for conflicts. When a conflict is detected, lower-confidence IDs are replaced with higher-ranked alternatives from the rank-5 list of the corresponding trajectory. This substitution is applied recursively until no conflicts remain in the global ID space.

**Algorithm 1:** Face-View ID Assignment

**Input:** Facial trajectories from three cameras, each trajectory with Rank-5 ID predictions and corresponding confidence scores

**Output:** Assigned identity for each trajectory without conflict

Concatenate trajectories from view 1, 2, 3 into an ordered list $\mathcal{T}$;
**foreach** *trajectory* $t \in \mathcal{T}$ **do**
    $t$.ID $\leftarrow$ $t$.Rank5[1];
    $t$.Conf $\leftarrow$ $t$.Confidences[1];
**repeat**
    Build mapping $\mathcal{M}$ from ID to set of trajectories sharing that ID;
    **foreach** *ID* $i$ *where* $|\mathcal{M}[i]| > 1$ **do**
        Sort $\mathcal{M}[i]$ by confidence in descending order;
        Let $t^* \leftarrow$ trajectory with the highest confidence;
        **foreach** *trajectory* $t \in \mathcal{M}[i] \setminus \{t^*\}$ **do**
            $replaced \leftarrow false$;
            **for** $k \leftarrow 2$ **to** 5 **do**
                $c \leftarrow t$.Rank5[$k$];
                **if** $c \notin \mathcal{M}$ ***or*** $\mathcal{M}[c] = \emptyset$ **then**
                    $t$.ID $\leftarrow c$;
                    $t$.Conf $\leftarrow t$.Confidences[$k$];
                    $replaced \leftarrow true$;
                    **break**;
            **if** *not replaced* **then**
                $t$.ID $\leftarrow -1$;
                $t$.Conf $\leftarrow 0$;
**until** *no ID conflict exists in* $\mathcal{T}$;
**return** *Assigned IDs for all trajectories in* $\mathcal{T}$;

If an unresolved conflict persists after all substitutions, the trajectory with the lowest confidence is assigned an ID of $-1$, indicating that the model failed to provide a reliable identity prediction.

#### 2.2.4. Cross-View ID Update

The Cross-View ID Update module fuses identity information from action and face views to achieve consistent identity association across cameras. This process relies on two priors derived from barn-level knowledge:

- Positional and numerical consistency. The number of individuals visible in the face-view cameras should correspond to the number of cattle present in the spatially overlapping region of the action-view cameras. Identity matching is therefore achieved by aligning spatial positions across views. However, because action-view cameras cover a larger area, cattle outside the face-view region are excluded from this process.
- Trajectory and temporal stability. During feeding, the cattle that can be reliably recognized are those remaining stationary at specific stalls for an extended duration. Thus, only the individuals successfully tracked and identified by the KBCFR model are considered valid candidates for identity assignment.

The procedure begins by projecting all tracked cattle from the action-view cameras onto the BEV map, making their global positions available to the face-view subsystem. We then identify which of these projected positions correspond to individuals actively feeding,

i.e., those visible within the face-view cameras. For recognized individuals, the center point of each face is projected onto the BEV grid representing the corresponding feeding stall. As shown in Figure 6, green pentagrams mark the BEV positions of cattle identified in the face views, while red and blue dots indicate detections from Action-View 1 and 2, respectively.

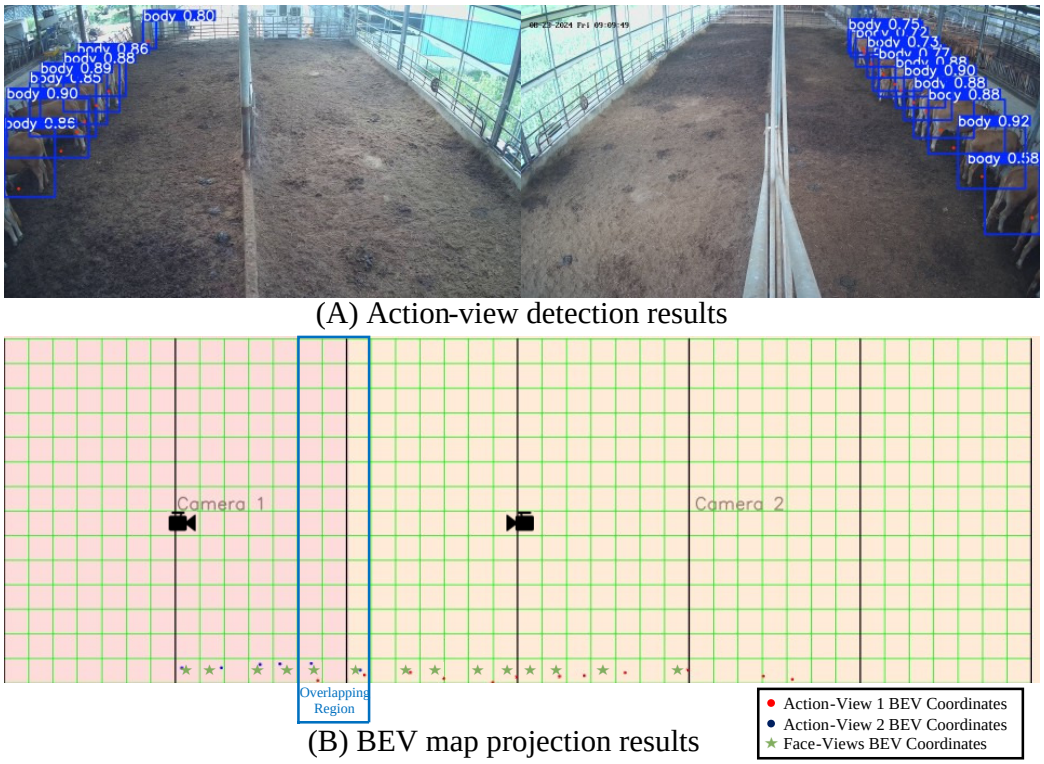



**Figure 6.** Body detection and BEV projection under grid mapping. The upper part shows detection results from Camera 1 and Camera 2, where bounding boxes are drawn for each detected cattle. The lower part illustrates the corresponding BEV projection of detected center points onto a grid-based spatial layout. Each dot represents a projected location of a detected instance.

To associate face IDs with tracking IDs, we compute the Euclidean distance between each face-view BEV point and the action-view BEV points. This distance-based strategy ensures that every recognized face is linked to a corresponding trajectory. Unlike cross-action matching, which requires IDs to fall into the same grid, this method reduces mismatches caused by projection errors that might otherwise prevent a face ID from being assigned to any trajectory.

Because face IDs are more reliable than action-view trajectories in the feeding area, we enforce that all trajectories within stall regions must be matched to a face ID. Trajectories without corresponding face matches, typically artifacts of over-detection by the tracker, are removed. Only individuals located within the BEV grids corresponding to feeding stalls are included in the cross-view identity update; those outside are excluded from the process.

## 3. Results

In this section, we present the evaluation results of our framework. We divide the analysis into three main components: detection, face recognition, and cross-view tracking. First, we assess the effectiveness of our YOLOv11x-pose-based detector, which is trained on a large-scale yet partially noisy dataset. We then evaluate the performance of the KBCFR-based face recognition module, which plays a critical role in periodically refreshing cattle identities across views. Finally, we analyze the performance of our cross-view tracking strategy, examining both qualitative and quantitative metrics to validate the framework's

ability to maintain identity consistency. Together, these results provide a comprehensive understanding of the framework's capabilities and limitations in long-term cross-view cattle tracking scenarios.

### *3.1. Detection*

We employ YOLOv11x-pose, the largest variant of the YOLO family, as the base detector. Despite noisy annotations in the training set, the detector maintains strong performance due to the large scale of the available data.

The training set comprised 193,963 images from 19 cattle. Specifically, 82,020 images were collected from Action-View 1 and 82,007 from Action-View 2 over a continuous 24-h monitoring period. An additional 29,936 images were obtained from the three Face-View cameras, of which 5742 were captured on the same day as the Action-View data. To reduce overfitting to these conditions, we included 24,194 additional Face-View images collected on different days under identical recording settings.

Annotations for Face-View images were manually refined and thus highly accurate. In contrast, Action-View annotations were more difficult to obtain. For each action segment, only the first and last frames were labeled manually, while intermediate frames were auto-generated using CVAT, introducing substantial label noise. To mitigate this, we retained only the manually annotated ground truth and discarded auto-generated labels. The number of training images was not reduced by this strategy because images and labels were not strictly one-to-one; as such, a single image may have simultaneous contained both refined and noisy labels. In extreme cases, even frames without labels were used during training as "negative samples", which helped improve generalization.

Although this approach introduces label sparsity, for example, cattle in motion may not be annotated in many frames, the large dataset size compensates for this limitation. As a result, the detector achieves strong performance despite partial supervision.

The training and validation curves of the detector are illustrated in Figure 7. The validation set contained 3000 images, including 1000 from the two Action-View cameras and 1000 from the three Face-View cameras. As shown in Figure 7, the detector performs well in bounding box prediction (metric B). The mAP50 consistently exceeds 85%, demonstrating robust localization, while the stricter mAP50:95 falls below 60% due to noisy annotations.

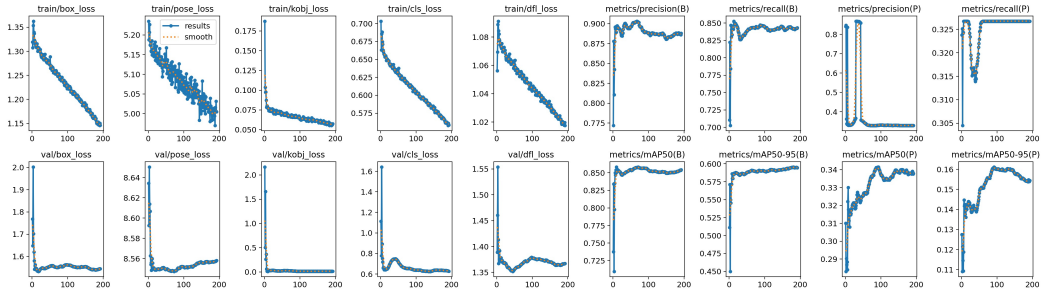



**Figure 7.** Training and validation curves of the detection and keypoint estimation model. (**Top**): Training losses and metrics. (**Bottom**): Validation losses and mAP scores across two tasks.

For keypoint detection (metric P), evaluation combines projected body center points from the Action-View cameras with facial landmarks from the Face-View cameras. Since these keypoints do not co-occur in the same frames, this metric provides limited interpretability. Nevertheless, the training and validation curves indicate that the model reached a stable, optimized state.

Figure 8 provides a qualitative comparison. In Figure 8A, where most cattle are walking, auto-generated labels show substantial noise: many bounding boxes and keypoints are misaligned. For instance, the cattle in the lower-left region has a bounding box far smaller

than its actual body. By contrast, Figure 8B shows the detector's predictions for the same frame, where most cattle are correctly identified and bounding boxes with center points are accurately estimated.

This result indicates that the detector remains effective and maintains strong performance even when trained on partially noisy annotations.

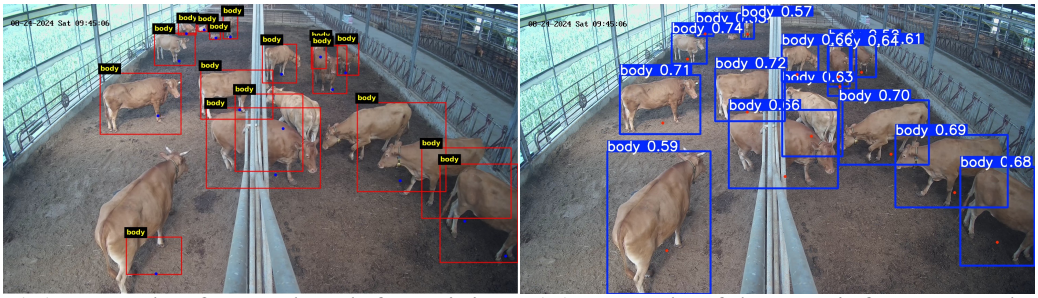



**Figure 8.** Comparison between ground truth annotations and detector predictions. (**A**) shows labeled bounding boxes and keypoints used for training. (**B**) shows the corresponding inference results with predicted boxes and confidence scores.

### *3.2. Face Recognition*

We adopt the KBCFR model from our prior work to perform cattle face recognition. Integrated as a module within the overall system, it receives face-view detection results from the detector as input. The model first tracks cattle faces across short temporal windows and then assigns an identity to each complete face trajectory.

Figure 9 presents the precision–recall curve of the face recognition module on the test set. Most cattle achieve an average precision (AP) greater than 90%, demonstrating that the module operates reliably under real-world barn conditions. This strong performance allows the system to periodically refresh identity assignments during feeding periods, thereby extending the continuity and stability of long-term tracking.

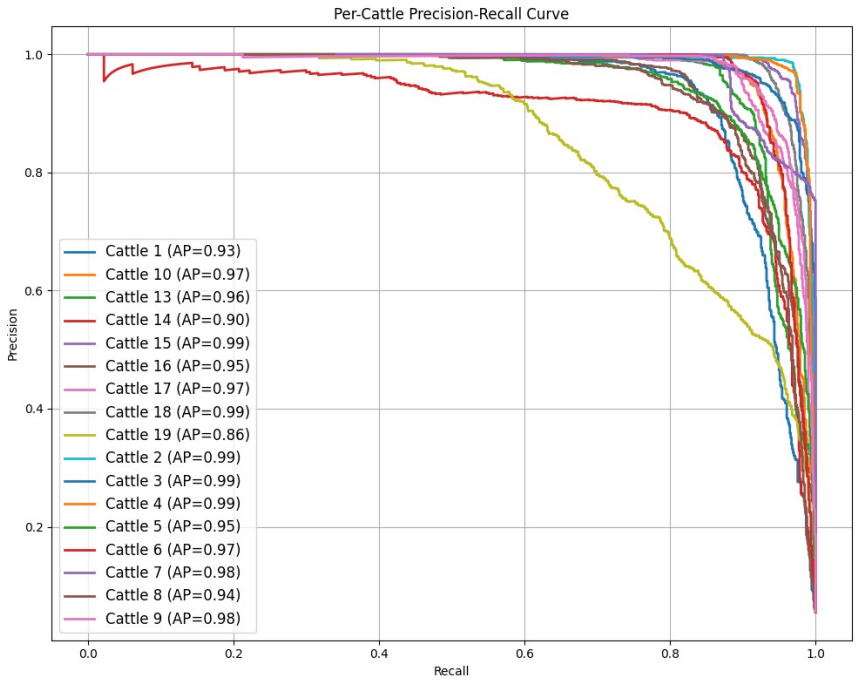



**Figure 9.** Precision–recall curves for individual cattle in face recognition. Each curve corresponds to one identity, with AP noted in the legend. The results show consistently high precision across most individuals.

It should be noted that Figure 9 does not include results for cattle 11 and 12. These two individuals are juveniles and are excluded from evaluation because of their significantly different facial appearance and behavioral patterns compared with adult cattle.

### *3.3. Cross-View Tracking*

For cross-view cattle tracking, we first perform independent single-camera tracking within each Action-View using a standard tracking-by-detection approach. The resulting trajectories are then matched across views using the proposed framework. When cattle faces are detected in the Face-View cameras, the corresponding face IDs are used to update and refine body tracking trajectories. In the absence of face detections, trajectory association is performed directly between the two Action-View cameras.

The framework is modular and extensible, allowing integration of various single-camera tracking algorithms. In this work, we adopt ByteTrack, which has demonstrated strong robustness to occlusion, as the default tracker for the individual Action-View cameras.

#### 3.3.1. Qualitative Results

Figure 10 presents a qualitative example. Figure 10A,B show tracking results from the two Action-View cameras, where most cattle are successfully tracked despite significant occlusions. Figure 10C–E show the corresponding face recognition results from the three Face-View cameras. Specifically, Figure 10C corresponds to the feeding area in Figure 10B, while Figure 10D,E correspond to the feeding region in Figure 10A. Face recognition IDs are correctly propagated back to the body trajectories, enabling identity updates across views. Even in overlapping areas, cattle 1 and 2 are correctly matched.

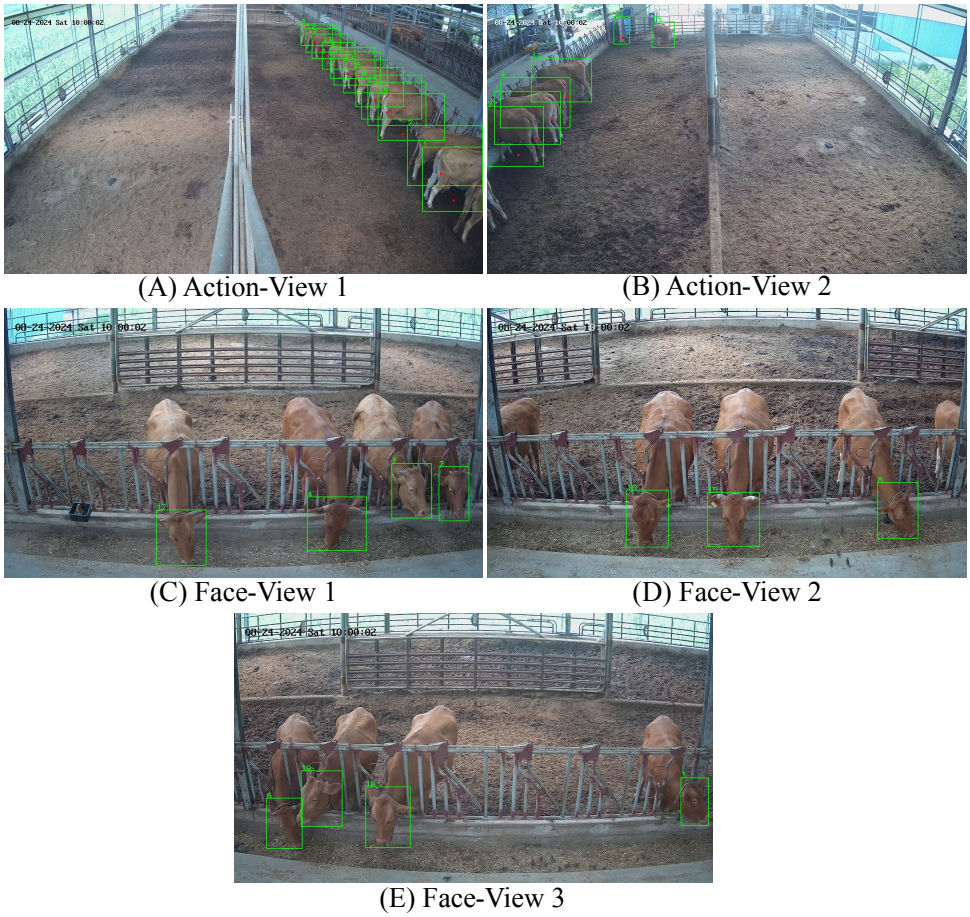
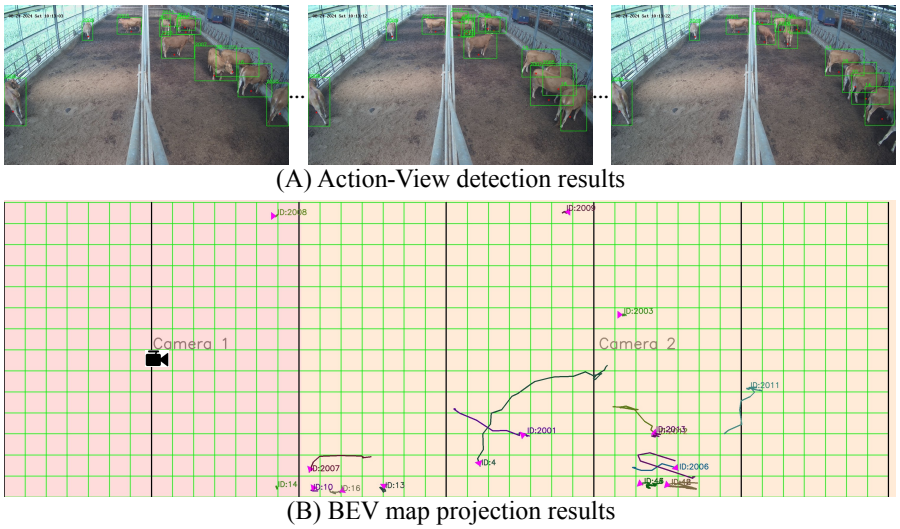



**Figure 10.** Example of cross-view tracking. (**A**,**B**) show tracking results in the two Action-View cameras. (**C**–**E**) show corresponding face recognition results from the Face-View cameras.

Two main types of failure cases are also observed:

- False detections. In Figure 10A, false positives occur in distant regions where bounding box predictions are unreliable. These errors lead to incorrect associations, as seen with cattle 6, 19, 10, and 3, which are mismatched with face IDs from Face-View 3.
- Duplicate IDs. In Figure 10B, a duplicate ID appears. One animal's trajectory is updated to ID 17 through face recognition, while another cattle outside the feeding area retains the original Action-View ID 17, resulting in a temporary conflict.

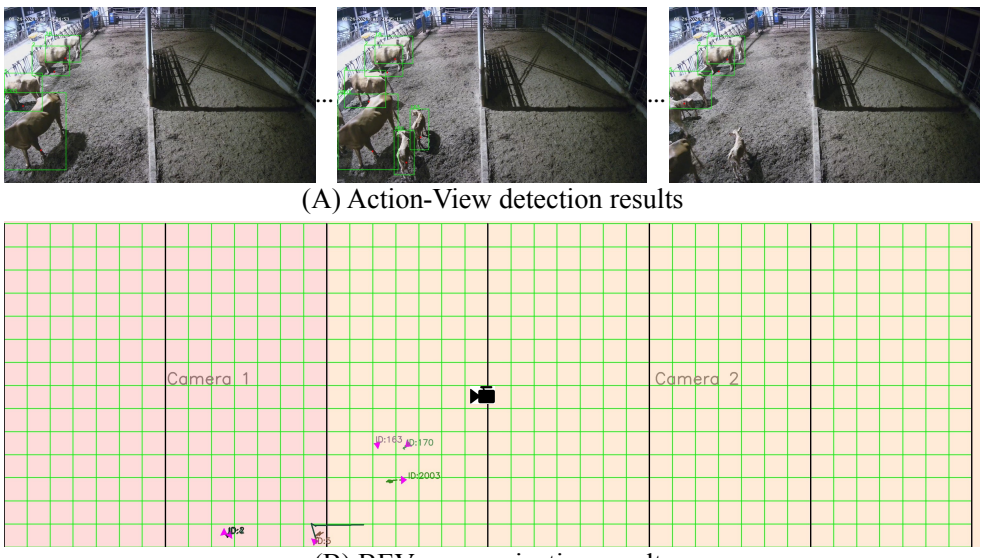
The first case can be mitigated by improving detector performance and increasing the number of cameras to enhance object detection and center point projections. The second case can be addressed by initializing Action-View trajectories with IDs distinct from those used in face recognition. For example, as shown in Figures 11 and 12, trajectory IDs are initialized starting from 2001 and updated only when a face-view match occurs. Although this inevitably introduces an ID switch, it enables more accurate identification during feeding.




(A) Action-View detection results

(B) BEV map projection results

**Figure 11.** Daytime feeding scenario. (**A**) Tracking results from Action-View 1, with bounding boxes and IDs. (**B**) Corresponding BEV trajectories.




(A) Action-View detection results

(B) BEV map projection results

**Figure 12.** Nighttime feeding scenario. (**A**) Tracking results from Action-View 2. (**B**) Corresponding BEV trajectories.

In the daytime feeding scenario, in Figure 11A, Action-View 1 produces bounding boxes (green) and IDs for each detected animal. In Figure 11B, ID 2007 is recognized as ID 10 after entering the face recognition zone. ID 16 frequently moves in and out of the feeding area and is sometimes tracked as ID 2005, but it is consistently corrected to ID 16 once face recognition occurs. Most of the remaining cattle retain IDs above 2000, as they never appear in the face recognition zone.

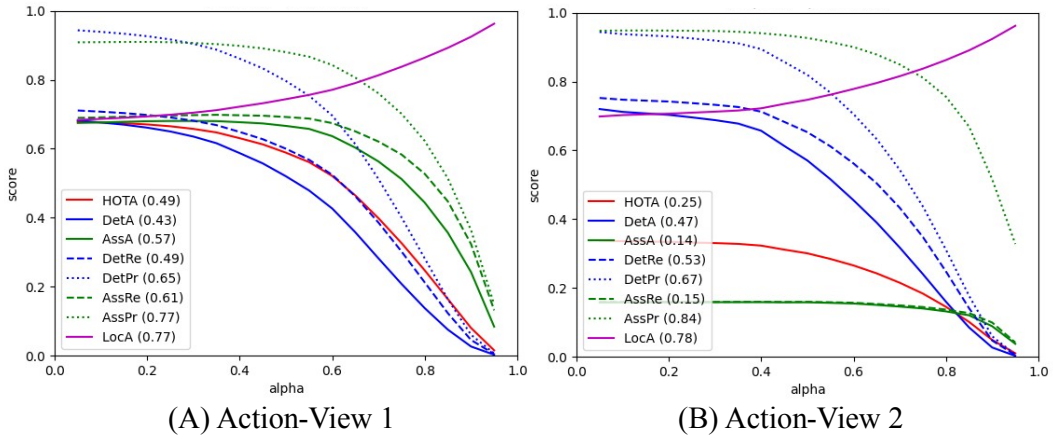
In the nighttime feeding scenario, Figure 12A shows the tracking results from Action-View 2, while Figure 12B maps the trajectories to the BEV. Due to reduced image quality at night, the action-view tracking exhibits more errors, such as the emergence of high-value IDs like 163 and 170. Nevertheless, the system still successfully updates identities within the feeding zone, demonstrating the robustness of the proposed method under varying lighting conditions.

Despite these occasional errors, a significant number of cattle trajectories are successfully updated with correct IDs, which represents a meaningful improvement for long-term tracking. In dense and occlusion-heavy environments like cattle barns, where motion patterns are irregular and ID switches are frequent, periodically refreshing identities using face recognition provides an effective mechanism to maintain identity consistency over time.

### 3.3.2. Quantitative Results

The quantitative results of the proposed framework are shown in Figure 13. Here, the variable *alpha* represents the matching threshold: the horizontal axis indicates the strictness of the criteria, while the vertical axis shows the corresponding performance scores.

The Higher Order Tracking Accuracy (HOTA) [33] metric evaluates the overall balance between detection and association performance. Detection Accuracy (DetA) measures how well the detected bounding boxes cover the true targets, while Association Accuracy (AssA) reflects the model's ability to maintain consistent target identities over time. Among the sub-metrics, Detection Recall (DetRe) indicates the proportion of true targets that are successfully detected, and Detection Precision (DetPr) denotes the proportion of correct detections among all predicted bounding boxes. Association Recall (AssRe) measures the proportion of frames in which a target is correctly assigned its identity, whereas Association Precision (AssPr) assesses whether each identity is exclusively assigned to a single target. Finally, Localization Accuracy (LocA) quantifies the spatial overlap between predicted bounding boxes and ground truth locations.




(A) Action-View 1 (B) Action-View 2

**Figure 13.** HOTA evaluation curves on two Action-View cameras. (**A**,**B**) show Detection, Association, and Localization Accuracy across different *alpha* values for Action-View 1 and Action-View 2, respectively.

When evaluating tracking, exact ID matching with ground truth is unnecessary. Instead, ground truth IDs are used to assess whether predicted IDs undergo misses or switches. It is sufficient to associate predicted IDs with ground truth IDs in the initial frame based on the bounding box.

Dataset -Specific Considerations

Two characteristics of our dataset affect the applicability of standard metrics:

- Noisy bounding boxes. Because many bounding box annotations contain errors, IoU-based detection metrics do not accurately reflect model performance. In contrast, ID labels are comparatively clean and provide a more reliable measure of tracking accuracy. For this reason, we focus on ID-based metrics such as AssPr, LocA, and AssA, while treating detection-related metrics as supplementary references.
- Cross-View ID Updates. During feeding, body tracking IDs are replaced with face IDs to ensure accurate identity assignment. However, HOTA-based metrics evaluate only the consistency of the initial tracking IDs and do not account for improvements from identity recognition. As a result, these updates are incorrectly counted as ID switches, which artificially lowers metric scores. Consequently, performance values reported during feeding periods may underestimate the true effectiveness of our method.

Action-View Comparison

Overall, Action-View 1 exhibits significantly better association performance compared to Action-View 2. This is because, during cross-view matching, we prioritize retaining trajectory IDs from Action-View 1. In cases of ID conflicts between the two views, the ID in Action-View 2 is updated to resolve the conflict. While this strategy preserves consistency in Action-View 1, it introduces a higher risk of identity fragmentation in Action-View 2.

On the other hand, Action-View 2 achieves higher Detection Accuracy, which may be attributed to its smaller coverage area and lower cattle density. In this region, cattle move less frequently and occlusion is reduced, allowing for more stable and precise detections.

Table 2 compares tracking performance across different configurations, where we replaced the single-camera tracker and the cross-camera matching module. To ensure fairness, all methods were integrated into the same framework, with only the corresponding modules substituted. Specifically, we compared ByteTrack and BoT-SORT (Ultralytics v8.3.3) [34] as single-camera backbones and evaluated RNMF [11] against our proposed BEV-based location matching algorithm for cross-camera association.

Based on the above considerations, we emphasize ID-based metrics (AssPr, AssA, LocA) as the most reliable indicators of performance, while HOTA and DetPr are included for reference.

**Table 2.** Tracking performance comparison of different methods under Action-View 1 and Action-View 2.

| Method | AssPr | LocA | AssA | HOTA | DetPr |
|---|---|---|---|---|---|
| BoT-SORT + RNMF in View 1 | 63.748 | 78.794 | 51.868 | 44.812 | 70.545 |
| BoT-SORT + RNMF in View 2 | 15.407 | 78.782 | 5.212 | 14.938 | 69.923 |
| ByteTrack + RNMF in View 1 | 57.539 | 77.186 | 45.279 | 43.180 | 65.068 |
| ByteTrack + RNMF in View 2 | 16.731 | 77.732 | 5.8811 | 16.181 | 66.630 |
| BoT-SORT + Ours in View 1 | 76.775 | 78.836 | 59.592 | 48.159 | 70.765 |
| BoT-SORT + Ours in View 2 | 82.092 | 78.809 | 12.920 | 23.424 | 70.037 |
| ByteTrack + Ours in View 1 | 77.108 | 77.253 | 57.186 | 48.839 | 65.412 |
| ByteTrack + Ours in View 2 | 84.481 | 77.787 | 14.284 | 25.078 | 66.886 |

As shown in Table 2, all methods achieve comparable Localization Accuracy. The moderate HOTA scores observed across all methods are partly due to the presence of noisy labels in the test set, which introduce uncertainty in detection and association evaluation.

In terms of association performance, both AssA and AssPr indicate that our BEV-based matching strategy outperforms RNMF. This suggests that precise spatial alignment in the BEV space leads to more accurate cross-camera association than trajectory-to-target matching in cluttered environments such as cattle barns.

Tracking metrics are generally consistent with the trends observed in Figure 13. Action-View 2 exhibits better Detection Precision, which is likely due to its smaller spatial coverage and lower cattle density, resulting in fewer false positives and improved detector performance.

### 3.3.3. Ablation Study

As part of the ablation study, Figure 14 illustrates how tracking performance varies over a full day for the two Action-View cameras. Performance is above average in the early morning when cattle activity is relatively low and tracking is less challenging.

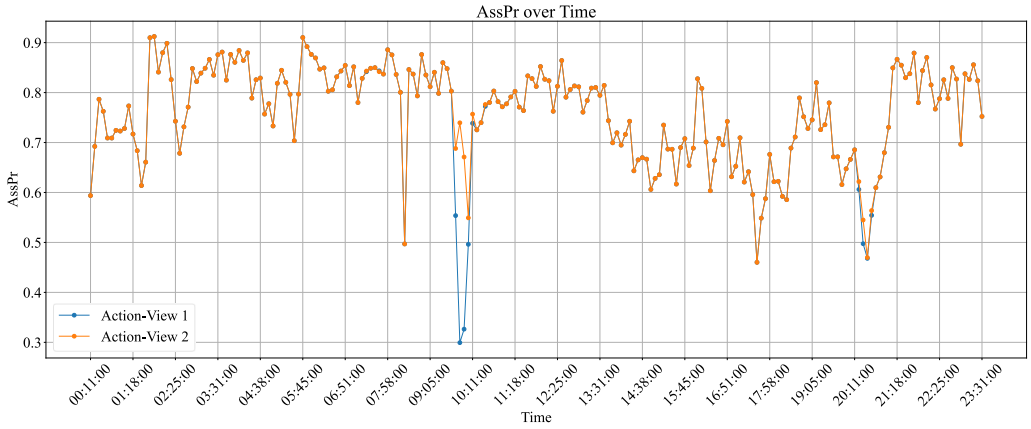



**Figure 14.** AssPr scores over time for Action-View 1 and Action-View 2. The plot shows per-interval tracking of performance variation across a full day.

A sharp performance drop occurs between approximately 9:00 and 10:00, corresponding to the first feeding period. During this time, increased activity and congestion, combined with Cross-View ID Updates being misinterpreted as ID switches, lead to a noticeable decline in performance. After the feeding session ends, performance gradually recovers.

This recovery is followed by a steady decline as cattle move to areas farther from the cameras, making detection more difficult. A second sharp drop is observed between 20:00 and 21:00, coinciding with the evening feeding session and mirroring the challenges of the morning peak.

To evaluate the impact of face recognition, we analyzed a 990-frame sequence captured during the feeding period under Action-View 1, comparing performance with and without the face recognition module. In this test, tracking IDs were initialized directly with face IDs, thereby avoiding early ID switch errors that might otherwise result from identity updates.

As shown in Figure 15, the model without face recognition performed slightly better during the first 90 frames. However, by around 360 frames, the face recognition module began to stabilize tracking, reducing ID inconsistencies. Over time, its benefit diminished, and the version without face recognition eventually achieved relatively better results. Nevertheless, both approaches exhibited declining performance over time, reflecting the major challenges posed by frequent movement and severe occlusion during feeding.

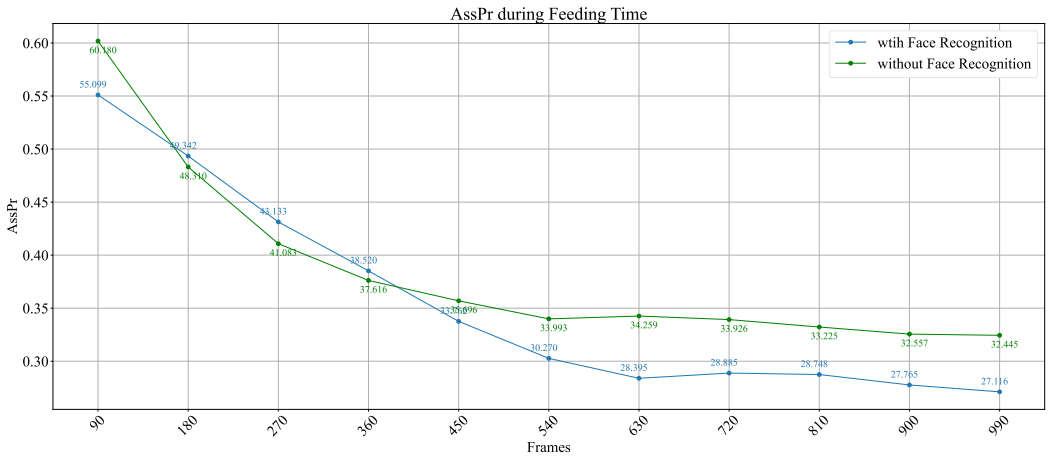



**Figure 15.** AssPr scores during the feeding period, highlighting the contribution of face recognition.

These findings suggest that cattle tracking performance follows clear temporal patterns linked to behavioral routines such as feeding. Understanding and adapting to these temporal dynamics is crucial for precision livestock farming, as it enables the design of more intelligent tracking systems and contributes to improved animal welfare monitoring.

## 4. Discussion

This work presents a unified framework for multi-object, multi-view cattle tracking in a real-world barn environment. The following discussion highlights the strengths of the proposed method, its current limitations, and potential directions to address these challenges.

### *4.1. Strengths*

The experimental results confirm that the proposed framework effectively captures both the spatial and temporal dynamics of cattle behavior. Unlike models that rely solely on appearance cues, our BEV-based spatial localization and trajectory matching strategy incorporates geometric constraints, enabling more reliable distinction between individuals with highly similar appearances. This capability is particularly valuable in barn environments, where cattle are densely packed and frequent occlusions occur.

A further strength is the framework's robustness to annotation noise. Despite the presence of imperfect labels in the training data, the system maintains stable tracking performance. This resilience suggests that the model can tolerate noisy supervision and generalize well to complex and variable barn conditions, which is essential for real-world deployment.

### *4.2. Limitations*

Despite these advantages, several limitations remain. First, the reliance on noisy annotations reduces the interpretability of precision-sensitive metrics. Errors in identity assignment can accumulate over long sequences, which undermines the reliability of detection-related metrics such as DetPr and HOTA. Moreover, cattle tracking differs fundamentally from conventional multi-object tracking benchmarks, which typically evaluate performance over short temporal segments. Our objective of continuous, day-long tracking introduces additional challenges, as identity drift and occlusion events accumulate over time. These challenges underscore the inherent difficulty of maintaining long-term identity consistency in barn environments.

A second limitation arises from the evaluation metrics themselves. Mainstream tracking metrics are primarily designed for short-term, open-world scenarios (e.g., pedestrian tracking at intersections) and thus have limited applicability to livestock monitoring. Two issues are particularly evident:

- Identity recognition is overlooked. Once an identity error occurs, subsequent evaluation of the affected trajectory becomes unreliable.
- Cross-camera re-identification (ReID) is not adequately captured. In conventional benchmarks, most objects do not reappear within the same camera's field of view, making these metrics unsuitable for evaluating long-term, multi-camera identity consistency.

These mismatches highlight the need for more appropriate evaluation methodologies that can capture both identity recognition and cross-camera association in livestock tracking scenarios. Developing such metrics will be an important direction for future research.

### *4.3. Potential Solutions*

Two complementary strategies could help mitigate these challenges. The first is to enhance annotation quality through additional manual labeling, which would improve supervision reliability but require significant time and labor. The second is to deploy additional cameras, thereby reducing the spatial coverage per view, increasing resolution, and minimizing occlusion. These hardware improvements, even without modifying the core framework, could substantially enhance the accuracy, robustness, and scalability of multi-camera cattle tracking in real-world barn environments.

## 5. Conclusions

This study proposed a cross-camera multi-object tracking framework for cattle monitoring in real-world barn environments, integrating multi-view spatio-temporal modeling, BEV-based trajectory matching, and face recognition-based identity updating to achieve robust and scalable long-term tracking. Extensive experiments on a large-scale multi-camera dataset with substantial annotation noise demonstrated strong localization and association performance, maintaining identity continuity under challenging conditions of high appearance similarity, crowding, and frequent occlusion. The main contributions include the development of a multi-object cross-camera tracking framework, the integration of BEV-based spatial matching with face recognition for reliable identity association, and comprehensive validation on a large, noisy, real-world dataset. Future work will focus on improving annotation quality and strategically deploying additional cameras to enhance resolution and reduce ambiguity, thereby further advancing the applicability of cross-camera cattle tracking systems in precision livestock farming.

**Author Contributions:** Conceptualization, S.H., A.F. and D.S.P.; methodology, S.H. and J.L.; software, S.H.; validation, S.H., A.F. and J.L.; formal analysis, S.H., A.F. and J.Y.; investigation, S.H. and Z.D.; resources, Y.J., S.Y. and D.S.P.; data curation, S.H. and J.P.; writing—original draft preparation, S.H. and Z.D.; writing—review and editing, A.F. and S.Y.; visualization, S.H.; supervision, J.Y., Y.J., S.Y. and D.S.P.; project administration, S.Y. and D.S.P.; funding acquisition, S.Y. and D.S.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00392406); 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).

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** The datasets used in this article are not readily available because this research has been carried out as part of an ongoing project funded by the Government of Korea. We have an agreement not to open the dataset until the project is concluded. Requests to access the data should be directed to Dong Sun Park, dspark@jbnu.ac.kr.

**Acknowledgments:** The authors gratefully acknowledge the support of the cattle farm located in Imsil, Republic of Korea, for permitting the deployment of surveillance cameras and facilitating data collection. This collaboration was essential for acquiring authentic livestock farming data in a real-world setting.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Xu, B.; Wang, W.; Guo, L.; Chen, G.; Wang, Y.; Zhang, W.; Li, Y. Evaluation of deep learning for automatic multi-view face detection in cattle. *Agriculture* **2021**, *11*, 1062. [CrossRef]
2. Myat Noe, S.; Zin, T.T.; Kobayashi, I.; Tin, P. Optimizing black cattle tracking in complex open ranch environments using YOLOv8 embedded multi-camera system. *Sci. Rep.* **2025**, *15*, 6820. [CrossRef] [PubMed]
3. Stanford, K.; Stitt, J.; Kellar, J.; McAllister, T. Traceability in cattle and small ruminants in Canada. *Rev. Sci.-Tech.-Off. Int. Epizoot.* **2001**, *20*, 510–522. [CrossRef] [PubMed]
4. Yang, L.; Liu, X.Y.; Kim, J.S. Cloud-based livestock monitoring system using RFID and blockchain technology. In Proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), New York, NY, USA, 1–3 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 240–245.
5. Klindtworth, M.; Wendl, G.; Klindtworth, K.; Pirkelmann, H. Electronic identification of cattle with injectable transponders. *Comput. Electron. Agric.* **1999**, *24*, 65–79. [CrossRef]
6. Wang, Z.; Fu, Z.; Chen, W.; Hu, J. A RFID-based traceability system for cattle breeding in China. In Proceedings of the 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), Taiyuan, China, 22–24 October 2010; IEEE: Piscataway, NJ, USA, 2010; Volume 2, p. 567.
7. Chen, X.; Yang, T.; Mai, K.; Liu, C.; Xiong, J.; Kuang, Y.; Gao, Y. Holstein cattle face re-identification unifying global and part feature deep network with attention mechanism. *Animals* **2022**, *12*, 1047. [CrossRef]
8. Nguyen, D.M.; Henschel, R.; Rosenhahn, B.; Sonntag, D.; Swoboda, P. Lmgp: Lifted multicut meets geometry projections for multi-camera multi-object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8866–8875.
9. Yamamoto, Y.; Akizawa, K.; Aou, S.; Taniguchi, Y. Entire-barn dairy cow tracking framework for multi-camera systems. *Comput. Electron. Agric.* **2025**, *229*, 109668. [CrossRef]
10. Salau, J.; Krieter, J. Instance segmentation with Mask R-CNN applied to loose-housed dairy cows in a multi-camera setting. *Animals* **2020**, *10*, 2402. [CrossRef]
11. He, Y.; Wei, X.; Hong, X.; Shi, W.; Gong, Y. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Trans. Image Process.* **2020**, *29*, 5191–5205. [CrossRef]
12. Sadeghian, A.; Alahi, A.; Savarese, S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 300–311.
13. Otsuka, K.; Mukawa, N. Multiview occlusion analysis for tracking densely populated objects based on 2-d visual angles. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, Washington, DC, USA, 27 June 2004–2 July 2004; IEEE: Piscataway, NJ, USA, 2004; Volume 1, p. I.
14. Nguyen, P.; Quach, K.G.; Duong, C.N.; Phung, S.L.; Le, N.; Luu, K. Multi-camera multi-object tracking on the move via single-stage global association approach. *Pattern Recognit.* **2024**, *152*, 110457. [CrossRef]
15. Wu, W.; Lai, J. Multi camera localization handover based on YOLO object detection algorithm in complex environments. *IEEE Access* **2024**, *12*, 15236–15250. [CrossRef]
16. Yang, C.Y.; Huang, H.W.; Kim, P.K.; Jiang, Z.; Kim, K.J.; Huang, C.I.; Du, H.; Hwang, J.N. An online approach and evaluation method for tracking people across cameras in extremely long video sequence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 16 2024; pp. 7037–7045.
17. Yoshida, R.; Okubo, J.; Fujii, J.; Amakata, M.; Yamashita, T. Overlap suppression clustering for offline multi-camera people tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 7153–7162.
18. Zheng, Z.; Qin, L. PrunedYOLO-Tracker: An efficient multi-cows basic behavior recognition and tracking technique. *Comput. Electron. Agric.* **2023**, *213*, 108172. [CrossRef]
19. Zhang, Y.; Sun, P.; Jiang, Y.; Yu, D.; Weng, F.; Yuan, Z.; Luo, P.; Liu, W.; Wang, X. Bytetrack: Multi-object tracking by associating every detection box. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 1–21.

20. Hao, W.; Ren, C.; Fan, Y.; Han, M.; Li, F. YSD-BPTrack: A multi-object tracking framework for calves in occluded environments. *Smart Agric. Technol.* **2025**, *11*, 100876. [CrossRef]
21. Guo, H.; Li, Z.; Ma, Q.; Zhu, D.; Su, W.; Wang, K.; Marinello, F. A bilateral symmetry based pose normalization framework applied to livestock body measurement in point clouds. *Comput. Electron. Agric.* **2019**, *160*, 59–70. [CrossRef]
22. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 107–122.
23. Zhang, Y.; Wang, C.; Wang, X.; Zeng, W.; Liu, W. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 3069–3087. [CrossRef]
24. Tu, S.; Lei, H.; Liang, Y.; Lyu, E.; Liu, H. RpTrack: Robust Pig Tracking with Irregular Movement Processing and Behavioral Statistics. *Agriculture* **2024**, *14*, 1158. [CrossRef]
25. Higaki, S.; Horihata, K.; Suzuki, C.; Sakurai, R.; Suda, T.; Yoshioka, K. Estrus detection using background image subtraction technique in tie-stalled cows. *Animals* **2021**, *11*, 1795. [CrossRef] [PubMed]
26. Fuentes, A.; Yoon, S.; Park, J.; Park, D.S. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. *Comput. Electron. Agric.* **2020**, *177*, 105627. [CrossRef]
27. Fuentes, A.; Han, S.; Nasir, M.F.; Park, J.; Yoon, S.; Park, D.S. Multiview monitoring of individual cattle behavior based on action recognition in closed barns using deep learning. *Animals* **2023**, *13*, 2020. [CrossRef] [PubMed]
28. Nasir, M.F.; Fuentes, A.; Han, S.; Liu, J.; Jeong, Y.; Yoon, S.; Park, D.S. Multi-camera fusion and bird-eye view location mapping for deep learning-based cattle behavior monitoring. *Artif. Intell. Agric.* **2025**, *15*, 724–743. [CrossRef]
29. Han, S.; Fuentes, A.; Park, J.; Yoon, S.; Yang, J.; Jeong, Y.; Park, D.S. Utilizing farm knowledge for indoor precision livestock farming: Time-domain adaptation of cattle face recognition. *Comput. Electron. Agric.* **2025**, *234*, 110301. [CrossRef]
30. Meng, Y.; Yoon, S.; Han, S.; Fuentes, A.; Park, J.; Jeong, Y.; Park, D.S. Improving known–Unknown cattle's face recognition for smart livestock farm management. *Animals* **2023**, *13*, 3588. [CrossRef]
31. Han, S.; Fuentes, A.; Yoon, S.; Jeong, Y.; Kim, H.; Park, D.S. Deep learning-based multi-cattle tracking in crowded livestock farming using video. *Comput. Electron. Agric.* **2023**, *212*, 108044. [CrossRef]
32. Jocher, G.; Qiu, J. *Ultralytics YOLO11*; Ultralytics Inc.: Frederick, MD, USA; London, UK; Madrid, Spain; Shenzhen, China, 2024.
33. Luiten, J.; Osep, A.; Dendorfer, P.; Torr, P.; Geiger, A.; Leal-Taixé, L.; Leibe, B. Hota: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef]
34. Aharon, N.; Orfaig, R.; Bobrovsky, B.Z. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv* **2022**, arXiv:2206.14651.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.