

Original papers

HanwooReID: Multi-view cattle re-identification with pose-aware transformer enhancements

Jiaqi Liu ^{a,b}, Alvaro Fuentes ^{a,b}, Shujie Han ^{a,b}, Yongchae Jeong ^a, Sook Yoon ^c, Dong Sun Park ^{a,b}^a Department of Electronics Engineering, Jeonbuk National University, Jeonju, 54896, South Korea^b Core Research Institute of Intelligent Robots, Jeonbuk National University, Jeonju, 54896, South Korea^c Department of Computer Engineering, Mokpo National University, Muan, 58554, South Korea

ARTICLE INFO

Keywords:

Individual re-identification

Hanwoo cattle

Pose-guided heatmap encoder

Viewpoint-constrained retrieval

Precision livestock farming

ABSTRACT

Identifying individual Hanwoo cattle automatically is particularly challenging: Hanwoo cattle exhibit highly similar visual characteristics, and unlike pedestrian re-identification, cattle undergo dramatic changes in appearance and scale as their poses and camera viewpoints vary. To address these challenges, we introduce HanwooReID, a large-scale multi-view dataset collected on real farms, comprising 9,929 images of 31 cattle captured under diverse viewpoints, poses, and lighting conditions. Building on this dataset, we propose a transformer-based framework that integrates a pose-guided heatmap encoder (PHE) to focus attention on identity-relevant regions and a viewpoint-constrained retrieval (VCR) strategy that projects hoof keypoints onto a bird's-eye view (BEV) plane to estimate cattle orientation. This BEV-based orientation estimation effectively filters out gallery candidates with inconsistent viewpoints, substantially improving matching accuracy under severe pose and view variations. Extensive experiments on both closed-set and open-set protocols show that our method outperforms existing baselines, achieving up to 7.4% and 6.8% improvements in mean Average Precision (mAP), respectively, demonstrating its effectiveness for precision livestock farming.

1. Introduction

In modern livestock farming, daily monitoring of cattle herds is a critical component for ensuring animal health and improving farming efficiency. By continuously tracking the behavior, health status, and growth of individual cattle, farmers can promptly detect abnormalities and take appropriate interventions, thus reducing economic losses and optimizing management practices. Traditional cattle management methods primarily rely on manual tagging or electronic ear tags (Pretto et al., 2024; Chang et al., 2022; Simanungkalit et al., 2021; Awad, 2016) to distinguish individuals, but these methods suffer from issues such as tag loss, high maintenance costs, and potential interference with natural animal behavior.

With the rapid advancement of computer vision technologies in recent years, image-based individual re-identification (ReID) has emerged as a promising approach for automated cattle management. ReID enables cross-time and cross-camera identification and tracking of the same individual without relying on physical tags, significantly enhancing the level of intelligent management in cattle farms (Ye et al., 2024).

Particularly in real-world scenarios, where tracking continuity may be interrupted by occlusion or camera switching, ReID modules can re-associate lost targets and maintain effective monitoring.

Existing animal ReID methods can be broadly categorized into two types: local feature-based methods and global feature-based methods. Local feature-based approaches leverage fine-grained details such as facial patterns (Lin et al., 2024; Bergamini et al., 2018; Bergman et al., 2024) or nose prints (Kumar and Singh, 2020; Dulal et al., 2025) for individual distinction, achieving high accuracy under ideal conditions but requiring high-quality close-up images, which are difficult to obtain consistently in farm environments. In contrast, global feature-based methods (Moskvyak et al., 2021; Wang et al., 2021) model the overall appearance of individuals, making them more practical for handling diverse viewpoints and pose variations. However, global methods still face challenges when dealing with cross-camera appearance changes, significant pose variations, and high inter-individual similarity.

Hanwoo cattle pose unique challenges for ReID due to their subtle and indistinct texture patterns across viewpoints. Unlike dairy cows,

* Corresponding author at: Department of Electronics Engineering, Jeonbuk National University, Jeonju, 54896, South Korea.

** Corresponding author at: Department of Computer Engineering, Mokpo National University, Muan, 58554, South Korea.

E-mail addresses: ryugagi@jbnu.ac.kr (J. Liu), afuentes@jbnu.ac.kr (A. Fuentes), shujiejulie@jbnu.ac.kr (S. Han), ycjeong@jbnu.ac.kr (Y. Jeong), syoon@mokpo.ac.kr (S. Yoon), dsark@jbnu.ac.kr (D.S. Park).<https://doi.org/10.1016/j.compag.2025.111117>

Received 24 June 2025; Received in revised form 1 September 2025; Accepted 13 October 2025

Available online 18 October 2025

0168-1699/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

which often have pronounced markings that aid in identification, Hanwoo cattle lack clear discriminative textures. Moreover, their appearance changes substantially between viewpoints, including significant variations in aspect ratio, making it difficult to extract robust and consistent features for accurate cross-view identification. These challenges are further exemplified in our dataset, which exhibits large appearance variations and spatial changes across viewpoints for the same individual, as illustrated in Fig. 1.

To address these challenges in Hanwoo cattle, which exhibit highly similar appearances and only subtle differences in body shape, we construct HanwooReID, a large-scale multi-camera dataset collected from two real cattle farms. The dataset captures images of the same individuals across multiple cameras and viewpoints, reflecting realistic variations in pose, illumination, and background. During data acquisition, we adopt and refine a camera calibration method suited to barn environments. Instead of relying on dedicated calibration targets, our approach estimates both intrinsic and extrinsic parameters based on manual measurements of fixed structural landmarks within the scene. These efforts provide a solid basis for developing and evaluating ReID models in real-world farm environments.

Built upon this foundation, we propose a Transformer-based ReID framework that leverages both image appearance and pose information. Specifically, we generate 2D pose heatmaps from each input and integrate them with visual features through a dedicated pose-guided heatmap encoder (PHE). This module provides structured spatial cues and helps guide the model's attention, enhancing its ability to distinguish visually similar individuals across different poses.

However, in our dataset, cattle often appear in a wide range of poses and orientations. As a result, even images of the same individual can exhibit large visual differences across viewpoints. Forcing the model to associate such dissimilar images solely based on appearance features may be unreasonable and inefficient. To address this, we propose a simple yet effective viewpoint-constrained retrieval (VCR) strategy, which estimates the cattle's facing direction by projecting hoof keypoints onto a bird's-eye view (BEV) plane. During evaluation, VCR filters out gallery images whose orientation differs significantly from that of the query. This viewpoint-aware filtering mechanism significantly improves the overall accuracy of our framework under severe pose and view variations.

In summary, our key contributions are as follows:

- **A Multi-View Hanwoo Cattle ReID Dataset:** We present HanwooReID, a comprehensive multi-view dataset for Hanwoo cattle ReID, containing 9,929 verified images of 31 individual cattle. Collected in authentic barn environments, it captures significant appearance variations caused by diverse poses and camera viewpoints. This dataset serves as a valuable resource for developing and evaluating fine-grained livestock identification methods under practical farm conditions.
- **Pose-Guided Heatmap Encoder for Attention Refinement:** We design a pose-guided heatmap encoder (PHE) that utilizes 2D keypoints predicted by a pose estimation model. The encoded pose-aware representations help guide the Transformer's attention mechanism toward identity-relevant regions, improving robustness under complex body deformations.
- **Viewpoint-Constrained Retrieval for Orientation-Aware Matching:** We introduce a lightweight viewpoint-constrained retrieval (VCR) strategy that estimates cattle orientation from monocular images by projecting hoof keypoints onto the BEV plane. By filtering gallery candidates based on viewpoint consistency, VCR effectively boosts ReID performance under large pose and view variations.
- **Deployment-Friendly ReID Framework for Farm Environments:** Our proposed algorithm is well-suited for real-world deployment in cattle farms. While it operates on instance-level image inputs, preceding steps such as detection, segmentation, and BEV

transformation can be automated in barn environments, making the overall pipeline highly feasible for deployment in precision livestock farming.

2. Related work

2.1. Existing datasets for Cattle ReID

High-quality datasets are essential for advancing cattle ReID research. Existing datasets vary in their strategies for representing individual cattle, often shaped by the available visual cues and environmental conditions. Broadly, these datasets fall into two categories: those focusing on specific anatomical regions such as the face or muzzle, and those utilizing full-body images captured from multiple viewpoints. Each category offers distinct advantages; however, most existing datasets face challenges in generalizing to unconstrained, multi-view scenarios. This limitation is particularly evident for breeds like Hanwoo cattle, which lack distinctive coat patterns or markings.

Region-specific Cattle ReID datasets. Several works have developed cattle ReID datasets targeting specific anatomical regions, primarily the face or muzzle. Ahmed et al. (2024) created a muzzle-based dataset spanning multiple breeds by exploiting unique nose patterns. Bakhshayeshi et al. (2023) gathered face images to support facial recognition in real farm environments. Weng et al. (2022) created facial image datasets of cattle, which include face data captured from diverse angles using handheld cameras, featuring Simmental and Holstein cows. Coffman et al. (2024) introduced CattleFace-RGBT, an RGB-T (thermal) facial landmark dataset comprising 2,300 image pairs designed to assist cattle health monitoring. Li et al. (2024) released ICRWE (Initial Cattle face Recognition in Wild Environments), a large-scale dataset consisting of 9,816 high-resolution cattle face images from wild settings, annotated with facial attributes. While these region-specific datasets enable fine-grained feature extraction by emphasizing localized anatomical regions such as the face or muzzle, their utility in practical farm settings is limited. High-resolution, close-range imaging is often infeasible under unconstrained conditions involving freely moving animals. Moreover, the inconsistent visibility of these features across varying viewpoints and poses renders such methods susceptible to occlusion, pose variation, and other real-world complexities.

Full-body cattle ReID datasets. Other research efforts focus on full-body cattle images to capture more comprehensive visual information. Ong et al. (2023) introduced CattleEyeView, a top-view dataset designed for detection, pose estimation, and tracking within barns. Li et al. (2018, 2021) studied lightweight identification under fixed side-view setups. Large-scale datasets such as those by Yu et al. (2024) and Gao et al. (2021) primarily cover Holstein-Friesian cattle captured from overhead viewpoints in controlled environments. Moskvayak et al. (2021) presented a yak dataset with side-view pose annotations for ReID and behavioral studies. Despite their scale, these datasets often exhibit limited viewpoint diversity or depend heavily on distinctive breed-specific markings, thereby limiting their generalizability to visually homogeneous breeds such as Hanwoo cattle.

2.2. ReID under pose and viewpoint variations

Pose and viewpoint variations present substantial obstacles to object ReID by causing large discrepancies in visual appearance across instances of the same identity. To address these issues, recent studies have proposed solutions based on convolutional neural networks (CNNs), Transformers, and hybrid frameworks that leverage both paradigms.

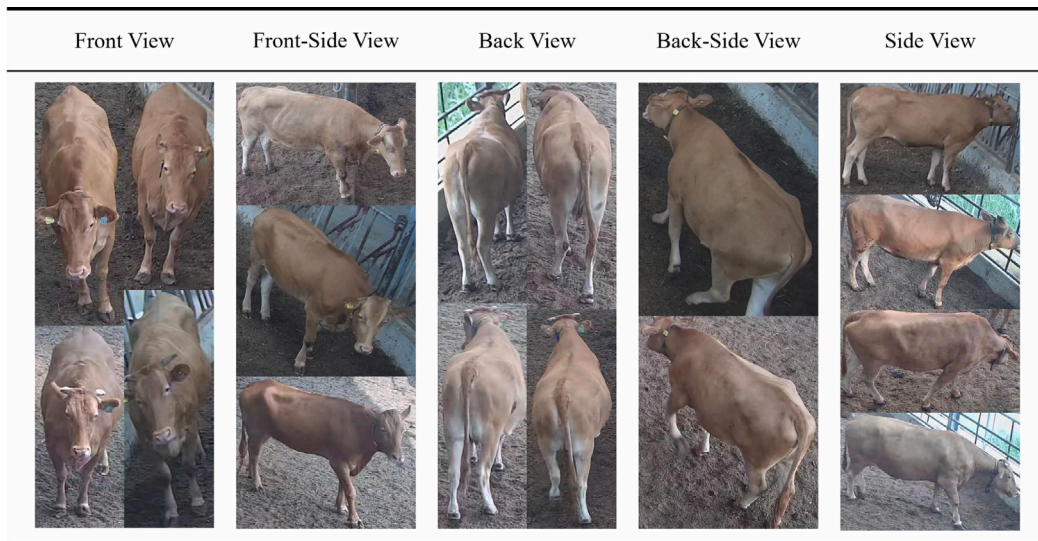


Fig. 1. Visualization of cattle images collected in our study. Images are captured from different viewpoints, leading to substantial changes in visual appearance across poses and angles. In addition to these viewpoint-induced variations, the spatial dimensions (aspect ratios and scales) of the instances also vary significantly. Unlike dairy cattle, Hanwoo cattle lack distinctive visual landmarks, making identification more challenging.

CNN-based and hybrid approaches. The AGW (Attention Generalized mean pooling with Weighted triplet loss) (Ye et al., 2021) enhances the BagTricks (Luo et al., 2019) framework through a non-local attention block, generalized-mean pooling, and weighted regularization triplet loss, improving robustness to cross-view variations in both single and cross-modality ReID. Classical metric learning methods such as Bruslund Haurum et al. (2020), Moskvayak et al. (2021) also leverage CNNs combined with triplet loss or handcrafted features to address appearance changes, though they may fall short under large pose variations. HAT (Hierarchical Aggregation Transformers) (Zhang et al., 2021) proposes a hybrid CNN-Transformer framework by aggregating multi-level CNN features via a deeply supervised aggregation module and calibrating them using Transformer-based global attention, preserving both semantic and fine-grained information for stronger representation learning. Nevertheless, the hybrid design of HAT still inherits the locality bias of CNNs, making it challenging to fully address the severe pose variations in our dataset, which motivates the introduction of pose-aware transformer designs.

Transformer-based approaches. Transformers have shown strong potential in modeling global dependencies for ReID. TransReID (He et al., 2021) introduces a Jigsaw patch module to handle local distortion and a side information embedding to reduce viewpoint bias. RotTrans (Rotation Invariant Transformer) (Chen et al., 2022) enhances robustness to spatial transformations by applying feature-level rotation and enforcing consistency through a rotation invariance regularization loss. PHA (Patch-wise High-frequency Augmentation) (Zhang et al., 2023) tackles the challenge of ViT's insensitivity to high-frequency details by introducing patch-wise high-frequency augmentation based on wavelet decomposition, coupled with a patch-wise contrastive loss, leading to enhanced high-frequency feature learning during training. PFD (Pose-guided Feature Disentangling) (Wang et al., 2022) focuses on occlusion and visibility issues by incorporating pose cues at both global and part levels, using pose-guided feature aggregation and pose-view matching modules within a Transformer framework to disentangle visible and occluded body parts effectively. These works illustrate the flexibility of Transformer-based pipelines in modeling complex spatial dependencies, which aligns well with the requirements of multi-view cattle ReID involving significant pose and viewpoint variations.

Pose and geometry-aware models. Pose-guided methods improve viewpoint invariance by leveraging pose information. PIE (Pose-Invariant

Embedding) (Zheng et al., 2019) uses pose box inputs with a multi-stream CNN for pose-invariant features. PETR (Position Embedding Transformation) (Liu et al., 2022, 2023) incorporates 3D positional embeddings with 2D features in Transformer architectures. Liu et al. (2018) synthesize new pose samples via a generator–guider–discriminator network to enhance training diversity. Junping et al. (2023) propose a Transformer-based pose-guided attention module that improves feature matching by reducing pose variation interference. Kishore et al. (2024) leverage pose estimation through heatmaps, key-points, and segmentation, integrating pose features into a two-stream Transformer-CNN architecture to enhance robustness to viewpoint changes. Xu et al. (2025) further integrate posture-aware temporal-smooth embeddings into a multi-camera cattle tracking framework under non-overhead views, enhancing robustness to appearance changes and occlusions while enabling accurate cross-camera association. These methods demonstrate that incorporating pose and geometric cues can effectively reduce viewpoint-induced appearance variation, which is particularly beneficial for multi-view cattle ReID with large pose deformation.

3. The proposed HanwooReID dataset

To support research on individual cattle ReID in realistic agricultural environments, we present HanwooReID, a multi-view dataset collected from real-world farms. The full dataset contains 134,071 images captured by multiple synchronized cameras, from which 9,929 cropped cattle instances were annotated with identity labels. These instance images were obtained through a series of filtering and sampling steps described in Sections 3.3.2 and 3.3.3, which remove severely occluded cattle, exclude instances with small bounding boxes, and apply action-specific frame sampling strategies.

To enable diverse evaluation protocols, we constructed 10 data sequences based on both open-set and closed-set settings. This section provides a detailed overview of the dataset, including the data collection setup, recording environments, annotation protocols, and camera calibration procedures.

3.1. Hanwoo farms

To ensure the diversity and realism of our dataset, we collected data from two distinct Hanwoo cattle farms. These farms differ in layout,

Table 1

Summary of the proposed dataset, including the number of cameras, images, recording duration, and annotated cows for each subset. The dataset covers both Imsil and Namwon farms, offering diverse multi-view data for evaluating ReID performance.

Dataset	Cameras	Cattle count	Image count	Duration
Imsil-day1	2	19	895	24 h
Imsil-day2	2	19	1456	24 h
Imsil-day3	2	19	2015	24 h
Imsil-day4	2	19	2724	24 h
Namwon-cam4	1	3	850	12 h
Namwon-cam5	1	9	1989	12 h

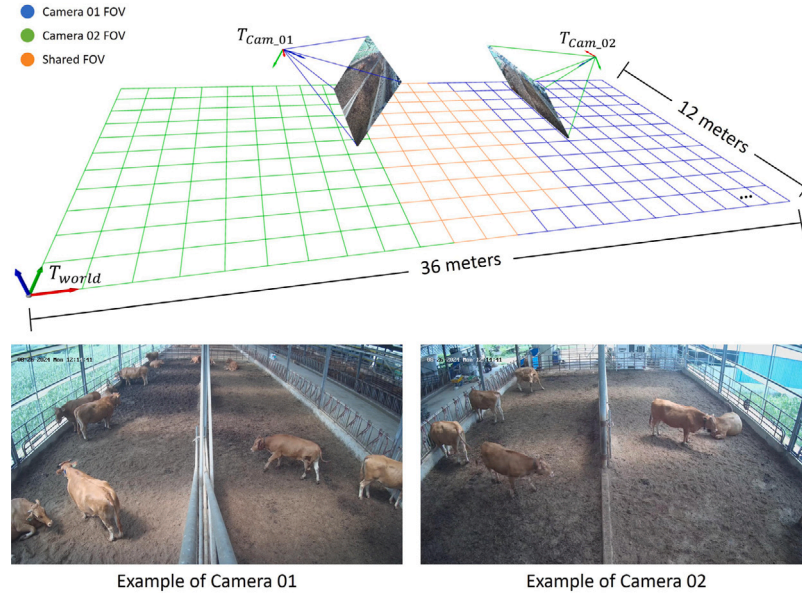


Fig. 2. A 3D view of the Imsil cattle farm (36 m × 12 m), showing two calibrated cameras installed to cover overlapping fields of view (FOVs). The blue and green grids represent the FOVs of Camera 01 and Camera 02, respectively, and the orange area indicates the overlapping region. Below, the images captured by both cameras are shown.

camera configurations, and herd composition, enabling the capture of a wide variety of real-world scenarios. Table 1 summarizes the key statistics of each dataset subset, including the number of cameras, cattle, images, and recording duration, providing a solid foundation for evaluating ReID methods across different conditions. In the following subsections, we provide a detailed description of each farm.

Imsil farm. Imsil Farm houses 17 adult Hanwoo cattle and 2 calves within an open barn area measuring 36 × 12 m. Two high-resolution Hikvision cameras were installed facing each other, enabling multi-view capture of the animals' daily activities. As illustrated in Fig. 2, the two cameras were carefully calibrated and positioned to cover overlapping fields of view (FOVs), with the blue and green grids representing the FOVs of Camera 01 and Camera 02, respectively. The orange area highlights the overlapping region, ensuring consistent visual tracking of cattle across views. Sample images from both cameras are also shown below the 3D layout for reference.

Data recording was conducted over four consecutive days at a frame rate of 60 Hz in RGB mode during daytime, ensuring consistent lighting conditions and high temporal resolution. The resulting dataset is divided into four subsets: Imsil-day1, Imsil-day2, Imsil-day3, and Imsil-day4, each containing 24 h of continuous footage. All subsets include the same 19 cattle individuals.

This dataset is especially valuable for studying fine-grained individual identification across varying viewpoints and poses, and provides a solid foundation for evaluating multi-view ReID performance in authentic farm environments.

Namwon farm. The Namwon dataset was collected from another real-world Hanwoo farm featuring a distinct spatial layout and camera

configuration compared to Imsil. As shown in Fig. 3, the farm consists of two spatially separated activity zones, each measuring approximately 10 × 10 m. A single Hikvision camera is installed in each zone, both oriented similarly toward the ground to effectively capture cattle activities within their respective areas.

The two zones house different cattle populations: one contains 9 adult Hanwoo individuals, while the other contains 3. Image sequences were continuously recorded for 12 h during the day under RGB mode. This setup introduces additional variability in scene composition, background clutter, and animal interactions, enriching the diversity of visual appearances captured.

In addition to the Imsil data, the Namwon dataset contributes further diversity by offering new environmental conditions and individual identities. Raw RGB images and annotations are provided in the same format as the Imsil dataset, enabling seamless integration for joint training, evaluation, or cross-scene analysis within the Hanwoo-ReID framework.

3.2. Data collection setup

To ensure the dataset reflects the visual and behavioral characteristics of Hanwoo cattle in realistic farm conditions, we designed a dedicated data acquisition system tailored to the challenges of livestock monitoring.

To obtain high-quality visual data, we used Hikvision cameras with a resolution of 3840 × 2160 and a frame rate of 60 Hz. This configuration enables accurate capture of cattle postures and movement trajectories over time. The combination of high resolution and high frame rate ensures clear representation of individual details, making

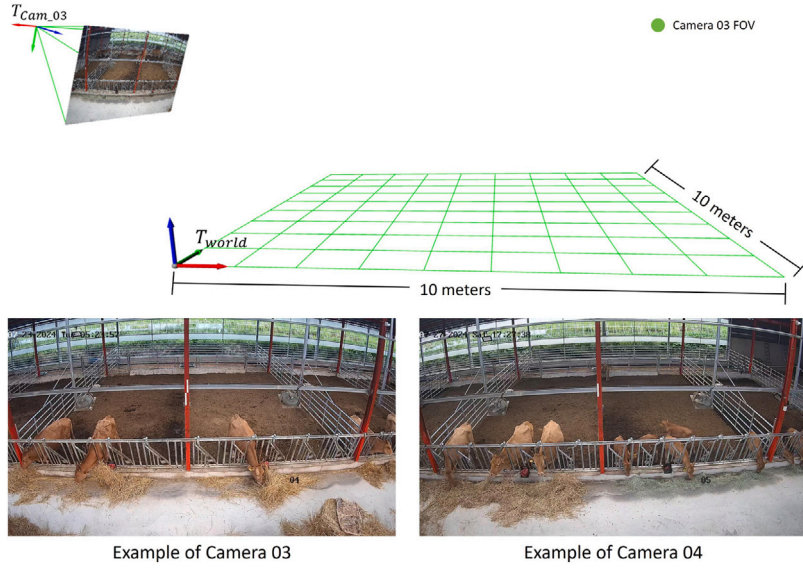


Fig. 3. A 3D view of the Namwon cattle farm, consisting of two spatially separated regions (each 10 m × 10 m), where two cameras are installed in different regions with similar orientation toward the ground. The layout and orientation of each camera are illustrated in the figure.

the dataset suitable for complex computer vision tasks such as cattle ReID.

To comprehensively capture the appearance variations of cattle under different poses, we continuously recorded their daily behaviors and movement trajectories from multiple viewpoints. This guarantees that each Hanwoo individual is thoroughly documented under multi-view and multi-temporal conditions, thereby reflecting their natural habits and dynamic characteristics. During daytime, when lighting conditions are favorable, the system records RGB color images; during nighttime or under low-light conditions, the cameras automatically switch to infrared (IR) mode, outputting grayscale images. This enables 24/7 data collection and enhances the dataset's spatiotemporal completeness and diversity.

Data collection was carried out at the previously described Imsil and Namwon farms, which differ in camera configurations and deployment strategies. These variations provide diverse visual conditions for evaluating model performance under multi-view scenarios.

3.3. Annotation and preprocessing

3.3.1. Camera calibration

Accurate geometric calibration is essential for multi-camera cattle monitoring systems. Building on prior efforts that leverage multi-viewpoint fusion and location mapping to enhance cattle behavior analysis in complex farm environments (Nasir et al., 2025), we perform geometric calibration for each camera to eliminate lens distortion and recover accurate visual representations of Hanwoo cattle.

Our calibration pipeline combines Direct Linear Transformation (DLT) with nonlinear optimization to estimate camera parameters. As illustrated in Fig. 4, given a set of 2D image points $\mathbf{x}_i = [u_i, v_i]^T$ and their corresponding 3D world coordinates $\mathbf{X}_i = [X_i, Y_i, Z_i]^T$, we estimate the intrinsic matrix $\mathbf{K}$, distortion coefficients $\mathbf{D}$, and extrinsic parameters $(\mathbf{R}, \mathbf{t})$. This process ensures that cattle positions and appearances are accurately projected across different camera views, facilitating reliable pose estimation and behavior recognition.

Normalization. We first normalize the image and world points to improve numerical stability. The 2D image points are normalized using a similarity transformation $\mathbf{N}_c$, and the 3D world points are normalized using a transformation $\mathbf{N}_w$:

$$\tilde{\mathbf{x}}_i = \mathbf{N}_c \mathbf{x}_i, \quad \tilde{\mathbf{X}}_i = \mathbf{N}_w \mathbf{X}_i \quad (1)$$

Initial DLT estimation. We estimate the initial projection matrix $\mathbf{P}$ via the Direct Linear Transformation (DLT) algorithm using the normalized coordinates. Given a set of normalized 3D-2D correspondences $\{(\tilde{\mathbf{X}}_i, \tilde{\mathbf{x}}_i)\}_{i=1}^n$, where $\tilde{\mathbf{X}}_i = [X_i, Y_i, Z_i, 1]^T$ and $\tilde{\mathbf{x}}_i = [u_i, v_i, 1]^T$, we aim to solve for the projection matrix $\mathbf{P} \in \mathbb{R}^{3 \times 4}$ such that:

$$\tilde{\mathbf{x}}_i \sim \mathbf{P} \tilde{\mathbf{X}}_i \quad (2)$$

To linearize this relation, each point correspondence generates two equations by rewriting the cross product $\tilde{\mathbf{x}}_i \times (\mathbf{P} \tilde{\mathbf{X}}_i) = \mathbf{0}$. This leads to a homogeneous linear system $\mathbf{A} \mathbf{p} = \mathbf{0}$, where $\mathbf{p}$ is the vectorized form of $\mathbf{P}$, and the coefficient matrix $\mathbf{A} \in \mathbb{R}^{2n \times 12}$ is constructed by stacking the contributions from each point:

$$\mathbf{A}_i = \begin{bmatrix} -\tilde{\mathbf{X}}_i^T & \mathbf{0}^T & u_i \tilde{\mathbf{X}}_i^T \\ \mathbf{0}^T & -\tilde{\mathbf{X}}_i^T & v_i \tilde{\mathbf{X}}_i^T \end{bmatrix} \quad (3)$$

We stack all $\mathbf{A}_i$ to form $\mathbf{A}$, and then solve for $\mathbf{p}$ using Singular Value Decomposition (SVD). Let $\mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T$, the solution $\mathbf{p}$ is given by the right singular vector corresponding to the smallest singular value, i.e., the last column of $\mathbf{V}$:

$$\mathbf{p} = \mathbf{v}_n \quad (4)$$

where $\mathbf{v}_n$ denotes the last column of matrix $\mathbf{V}$. Reshaping $\mathbf{p}$ gives the initial estimate of the projection matrix $\mathbf{P} \in \mathbb{R}^{3 \times 4}$.

Finally, the denormalized projection matrix $\mathbf{P}'$ is recovered by applying the inverse of the normalization transformations $\mathbf{N}_c$ and $\mathbf{N}_w$:

$$\mathbf{P}' = \mathbf{N}_c^{-1} \mathbf{P} \mathbf{N}_w \quad (5)$$

Nonlinear optimization. Based on the denormalized projection matrix $\mathbf{P}'$ obtained from the DLT and normalization reversal, we perform nonlinear optimization to refine the camera parameters, including the intrinsic matrix $\mathbf{K}$, distortion coefficients $\mathbf{D} = [k_1, k_2, p_1, p_2, k_3]^T$, and extrinsic parameters $(\mathbf{R}, \mathbf{t})$. As a preliminary estimate, we extract the left 3×3 submatrix $\mathbf{M}$ of $\mathbf{P}'$ and perform QR decomposition on $\mathbf{M}^{-1}$ to obtain $\mathbf{R}$ and $\mathbf{K}$. The translation vector is then computed as $\mathbf{t} = \mathbf{K}^{-1} \mathbf{P}'_4$, where $\mathbf{P}'_4$ denotes the fourth column of $\mathbf{P}'$. The parameters are subsequently refined by minimizing the total reprojection error between the observed 2D points and the projections of the corresponding 3D points.

Let $\hat{\mathbf{x}}_i$ denote the reprojected 2D point of the i th 3D world point $\mathbf{X}_i$, computed via a projection function $\pi(\cdot)$ that accounts for perspective projection as well as radial and tangential distortion:

$$\hat{\mathbf{x}}_i = \pi(\mathbf{K}, \mathbf{D}, \mathbf{R}, \mathbf{t}, \mathbf{X}_i) \quad (6)$$



Fig. 4. Camera calibration results from two different viewpoints. Traditional calibration typically relies on detecting keypoints from a printed chessboard or calibration board. However, in this scene, although the board is visible, its distance from the camera limits its coverage and resolution, leading to inaccurate parameter estimation. To address this, we manually selected 2D keypoints as ground truth (blue solid circles) and estimated intrinsic and extrinsic parameters using DLT followed by Levenberg–Marquardt optimization. The reprojected points (red hollow circles) demonstrate the accuracy of the estimated parameters.

The optimization objective is defined as:

$$\min_{\mathbf{K}, \mathbf{D}, \mathbf{R}, \mathbf{t}} \sum_i \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \quad (7)$$

where $\mathbf{x}_i$ is the observed 2D point.

We adopt an iterative nonlinear least squares method to jointly optimize all parameters, initialized with values obtained from the decomposition of the projection matrix. This refinement enhances the geometric accuracy of the camera model and ensures a more precise correspondence between 3D world coordinates and their 2D image projections. The visualized calibration results can be seen in Figs. 2 and 3.

3.3.2. Object annotations

To support the behavior-aware sampling strategy described in Section 3.3.3, we annotated each individual's action, bounding box, identity label, and camera ID. These annotations enable more effective selection of representative frames for model training by capturing diverse motion characteristics. Using a CVAT-based tool (Anon, 2025), trained annotators labeled each individual across 12 predefined action classes and assigned consistent identity labels across all camera views, ensuring both spatial and temporal coherence throughout the scene.

To improve labeling consistency, annotators referenced prior work on cattle behavior understanding. In particular, the study by Fuentes et al. (2020) demonstrated the effectiveness of deep learning-based hierarchical recognition frameworks for identifying both individual and group cattle activities using spatio-temporal cues. Their approach utilized RGB video data collected across various lighting conditions and locations, revealing the diversity and complexity of cattle behaviors. Inspired by this, our annotation process also emphasized visual context and temporal consistency across frames and camera views.

For visually similar actions such as standing and walking, annotators leveraged all available camera angles and specific visual cues to disambiguate and accurately classify each behavior. In overlapping regions where cows were severely occluded, we assigned the same action label as that of the partially visible portion, provided sufficient

evidence was available from other views. A summary of all annotated action classes is provided in Table 2.

Bounding boxes were initially generated using an object detection model. To assess occlusion severity, we computed the Intersection over Union (IoU) between each cow's bounding box and all other bounding boxes within the same image. If the maximum IoU with any other object exceeded a certain threshold of 0.5, indicating substantial overlap, the instance was considered heavily occluded and excluded from the dataset. The IoU is defined as:

$$\text{IoU} = \frac{|B_i \cap B_j|}{|B_i \cup B_j|}, \quad (8)$$

where B_i and B_j denote the bounding boxes of two distinct instances in the same image, and $|\cdot|$ represents the area of a region.

Each instance image is annotated with a comprehensive set of metadata, including the unique ID, the associated action label, and image-specific attributes such as timestamp and resolution. In addition, we record detailed camera parameters for each image, including both intrinsic and extrinsic calibration values, to support precise geometric reasoning and potential multi-view analysis. Fig. 5 shows several representative annotated instances depicting common actions such as Standing, Walking, Eating, and Resting, along with the corresponding Hanwoo IDs and camera IDs, providing a visual overview of the annotation format and behavioral diversity in our dataset.

3.3.3. Behavior-aware sampling strategy

Based on the action annotations described in the previous subsection, we implement a sampling strategy tailored to the motion characteristics of each behavior. For actions involving rapid movement such as walking, standing up, or mounting, we sample approximately one instance image per minute for each individual. In contrast, for slow or static behaviors such as sleeping, resting, and eating, we retain only one image for each continuous action segment to avoid redundancy.

In addition to temporal constraints, we apply a spatial filtering step to ensure visual quality. Specifically, we retain only those instance images where the bounding box height and width exceed predefined thresholds, thereby excluding low-resolution or partially captured



Fig. 5. Representative annotated instance images showing various cattle actions (Standing, Walking, Eating, and Resting). Each image is labeled with the corresponding Hanwoo ID and camera ID.

Table 2
Summary of annotated action classes for Hanwoo cattle behavior.

Action label	Description
Walking	Cow is moving; feet are in the air or in motion across frames.
Standing	Cow is stationary and upright, not performing other actions.
Resting	Cow is lying down with legs tucked in, head up, eyes open.
Sleeping	Cow is lying with its head down on the body and eyes closed.
Eating	Cow's head is through the fence, consuming feed.
Standing Up	Transition from resting to standing.
Lying Down	Transition from standing to resting.
Self-Grooming	Cow is licking itself while standing or lying.
Fighting Head-butt	Cow uses its head to hit or collide with another cow.
Fighting Push	Cow pushes another using body, neck, or shoulder.
Social Licking (Active)	Cow actively licks another cow.
Social Licking (Passive)	Cow being licked by another cow.

samples. This strategy is consistently applied to both farms, ensuring uniformity and data quality across the dataset.

4. Method

Distinguishing visually similar Hanwoo cattle is particularly challenging due to their subtle appearance differences and uniform coat color. To enable the model to focus on identity-specific features beyond global appearance, we explicitly incorporate structural cues derived from anatomical keypoints. Our dataset includes full-body images captured from diverse viewpoints, enabling models to extract fine-grained identity cues without relying on localized annotations such as facial landmarks or texture patterns.

4.1. Framework overview

We propose a viewpoint-aware ReID framework based on the TransReID architecture, specifically designed for the challenges of multi-camera, cross-view cattle ReID. The overall design is illustrated in Fig. 6. Our approach enhances the baseline architecture by incorporating a pose-guided heatmap encoder (PHE) module and a viewpoint-constrained retrieval (VCR) strategy.

To extract structural cues from the input image, we employ the PHE module. Specifically, we first detect anatomical keypoints, such as eyes, shoulders, knees, and hooves, using a pretrained animal pose estimation model. Existing high-performance methods such as RTMPose (Jiang et al., 2023) and ViTPose (Xu et al., 2022) have demonstrated strong generalization across animal species and can be adapted

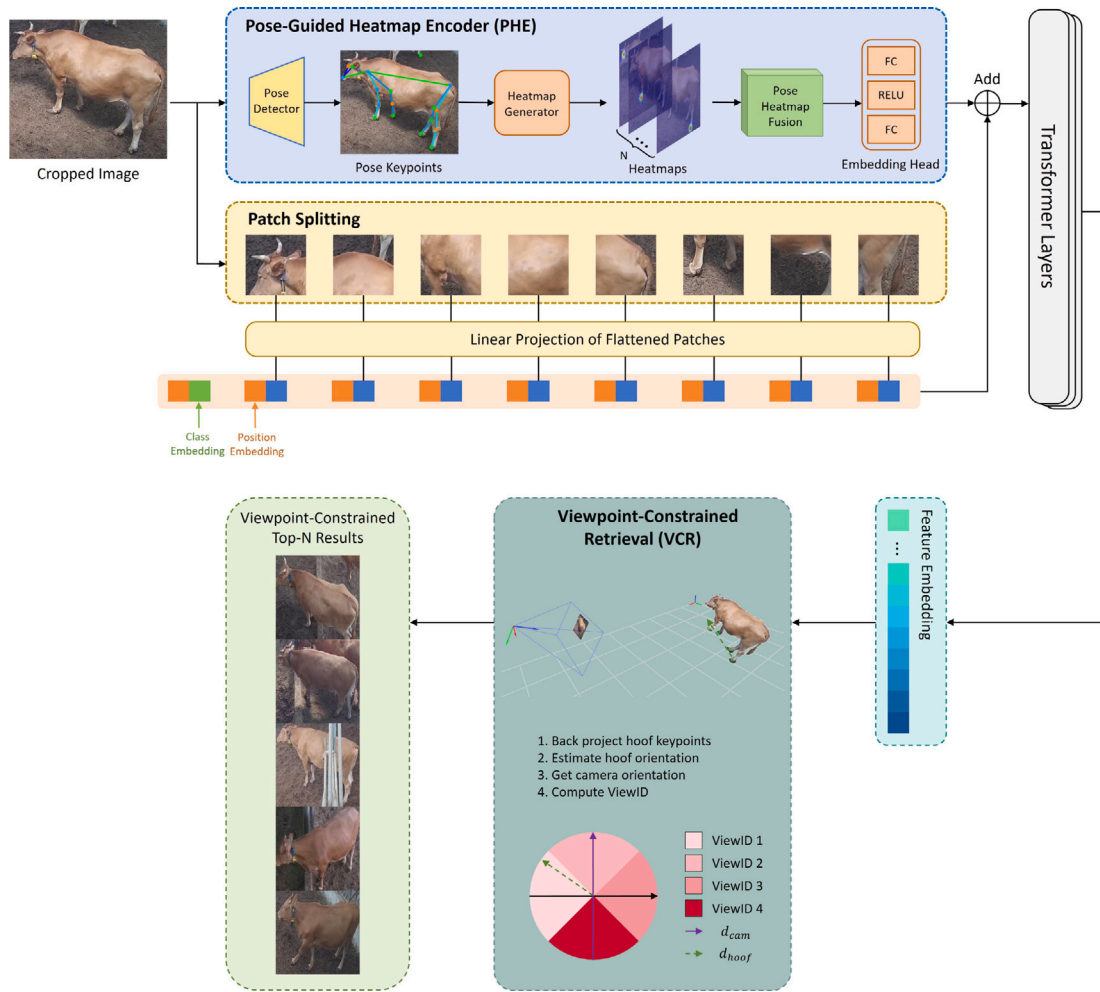


Fig. 6. Overview of our proposed framework. The PHE extracts 2D pose keypoints from the cropped input image, generates corresponding Gaussian heatmaps, and encodes them into pose feature maps. These pose features are fused with visual tokens from the image branch and processed by transformer layers for representation learning. The learned representations are fed into the VCR module, which predicts viewpoint labels and retrieves viewpoint-constrained top- N candidates from the gallery.

to our cattle-specific setting. These models are typically trained on publicly available animal pose datasets, including AP-10K (Yu et al., 2021) and Animal-Pose (Cao et al., 2019), which provide dense keypoint annotations across a wide range of quadruped species.

The predicted keypoints are converted into heatmaps that represent the spatial confidence distribution of each anatomical point. These heatmaps are then embedded through an embedding head and fused with visual tokens derived from patch embeddings. Specifically, the input image is first divided into fixed-size patches, which are flattened and projected linearly to form patch-level visual tokens. Learnable positional embeddings are added to retain spatial information, and an additional class embedding is prepended to represent the global context. By integrating pose information, the model is guided to attend to semantically meaningful body regions, thereby improving robustness to occlusion and severe viewpoint variations.

In addition to structural encoding, we address viewpoint discrepancy through VCR strategy. Based on the 2D pose keypoints of the four hooves, we estimate each individual's direction in the BEV coordinate system. The relative angle between the estimated cattle orientation and the camera's optical axis is used to assign a discrete ViewID. During evaluation, gallery images whose ViewIDs differ significantly from that of the query are excluded from the matching process. The VCR strategy effectively reduces visual ambiguity and improves recognition performance in complex multi-camera environments.

To train the proposed framework effectively, we adopt a combination of identity classification loss (ID Loss) (Zheng et al., 2017) and metric learning loss (Triplet Loss) (Liu et al., 2017). The ID Loss is implemented as a standard cross-entropy loss that encourages the network to correctly classify each individual cattle identity. The Triplet Loss, on the other hand, aims to increase the distance between features of different identities while pulling closer features of the same identity under different views and poses. Together, these two losses contribute to discriminative feature learning and retrieval accuracy, which are critical for cattle ReID under complex pose and view conditions.

Compared with the baseline, which encodes viewpoint information using ViewID Embedding (VID), a learnable embedding tied to each camera view identifier, our method, VCR, adopts a fundamentally different strategy by introducing an orientation-based matching constraint. Ablation studies indicate that while the VID has shown improvements in other domains, such as vehicle ReID, it does not consistently improve performance for cattle ReID and may occasionally reduce accuracy. In contrast, the proposed VCR approach utilizes estimated cattle orientation to filter out visually inconsistent matches during evaluation. This strategy can be seamlessly integrated into a broader cattle monitoring system, where geometric priors contribute to more accurate and robust ReID under diverse camera viewpoints.

In summary, our framework combines a PHE module and a VCR strategy to enhance cross-view cattle ReID. Detailed descriptions of the modules follow in the next subsections.

4.2. Pose-guided heatmap encoder

To enhance the model's ability to capture the anatomical structure and viewpoint-dependent appearance of Hanwoo cattle, we propose a PHE that transforms 2D pose keypoints into spatial representations, which are subsequently used as position encodings.

For each cropped cattle image, we extract anatomical keypoints using a pose estimation module embedded in our framework. The detected 2D keypoints are denoted as $\mathbf{K} \in \mathbb{R}^{J \times 3}$, where $J = 17$ is the number of keypoints and each row contains the (x, y, conf) values of a single keypoint. These keypoints are then transformed into Gaussian heatmaps, resulting in a multi-channel pose heatmap tensor $\mathbf{H}_{\text{pose}} \in \mathbb{R}^{J \times H \times W}$, where each channel represents the spatial confidence distribution of one anatomical location.

To form the final input for position encoding, we explore multiple fusion strategies to aggregate the heatmaps across keypoints, including max pooling, average pooling (mean), and concatenation (no fusion). Empirically, we found that max pooling yields the best trade-off between preserving keypoint-specific information and reducing computational complexity. The max-pooled heatmap is computed as:

$$\mathbf{H}_{\text{agg}} = \max_{j=1, \dots, J} \mathbf{H}_{\text{pose}}^j \in \mathbb{R}^{1 \times H \times W} \quad (9)$$

This aggregated heatmap $\mathbf{H}_{\text{agg}}$ is then passed through a dedicated convolutional encoder $\mathcal{E}_{\text{pose}}$ composed of two consecutive 1×1 convolutional layers with ReLU activation:

$$\mathbf{E} = \mathcal{E}_{\text{pose}}(\mathbf{H}_{\text{agg}}) \in \mathbb{R}^{C \times H \times W} \quad (10)$$

where C is the desired embedding dimension. The resulting positional feature map $\mathbf{E}$ is flattened and added to the corresponding patch embeddings as an attention prior. This module enables the model to incorporate anatomical priors from pose keypoints in a purely 2D manner, guiding attention toward semantically meaningful regions.

We provide ablation studies in Section 5.4 to evaluate the impact of different fusion strategies on the overall ReID performance.

Algorithm 1: VCR Strategy

Input: Hoof keypoints $\mathbf{h}^{2D}$; camera parameters $(\mathbf{K}, \mathbf{R}, \mathbf{t})$; ranked candidate list C from feature embedding

Output: Viewpoint-constrained Top- N results

```

/* Viewpoint estimation */
1  $\mathbf{h}^{3D} \leftarrow \text{BackProject}(\mathbf{h}^{2D}, \mathbf{K}, \mathbf{R}, \mathbf{t})$ ;
2  $\mathbf{d}_{\text{hoof}} \leftarrow \text{EstimateHoofOrientation}(\mathbf{h}^{3D})$ ;
3  $\mathbf{d}_{\text{cam}} \leftarrow \text{GetCameraForwardVector}(\mathbf{R})$ ;
4  $\text{view} \leftarrow \text{MapAngleToView}(\mathbf{d}_{\text{cam}}, \mathbf{d}_{\text{hoof}})$ ;

/* Constrained retrieval on ranked candidates */
5  $C_{vc} \leftarrow \emptyset$ ;
6 foreach  $g \in C$  do
7   if  $\text{view}_g = \text{view}$  then
8      $C_{vc} \leftarrow C_{vc} \cup \{g\}$ ;
9 return Top- $N$  from  $C_{vc}$ ;

```

4.3. Viewpoint estimation and constrained retrieval

Cattle exhibit a wide range of orientations relative to the camera in open-field environments, resulting in significant appearance variations. These viewpoint-induced differences pose a major challenge for ReID models, especially when visually dissimilar images of the same individual are captured from different directions. Relying solely on image-level features is often insufficient and may require large-scale training data to achieve robust generalization. Instead of relying on the model to learn viewpoint-invariant embeddings, our approach explicitly estimates cattle orientation and uses this information to constrain the retrieval process through the VCR strategy, as detailed in Algorithm 1. This

strategy avoids comparisons between visually inconsistent views by filtering candidate matches according to their estimated orientations.

To estimate cattle orientation from a monocular image, we utilize four hoof keypoints predicted by a pose estimation model: left front (LF), right front (RF), left rear (LR), and right rear (RR). These keypoints, assumed to lie on the ground plane at fixed height $h = 0$, are represented in image coordinates $\mathbf{p}_i = (u_i, v_i)$, where $i \in \{\text{LF}, \text{RF}, \text{LR}, \text{RR}\}$. Given the camera intrinsic matrix $\mathbf{K}$ and extrinsic parameters $(\mathbf{R}, \mathbf{t})$, each keypoint is back-projected onto the ground plane by intersecting the corresponding camera ray with the ground, based on the camera pose. Formally, the back-projected point $\mathbf{q}_i$ in world coordinates can be computed as:

$$\mathbf{q}_i = \lambda_i \mathbf{R}^\top \mathbf{K}^{-1} \tilde{\mathbf{p}}_i + \mathbf{c}, \quad (11)$$

where $\tilde{\mathbf{p}}_i = (u_i, v_i, 1)^\top$ is the homogeneous image coordinate, $\mathbf{c} = -\mathbf{R}^\top \mathbf{t}$ denotes the camera center in world coordinates, and the scalar λ_i is chosen such that $\mathbf{q}_i$ lies on the ground plane $h = 0$.

The cattle's body orientation is defined as the vector from the midpoint of the rear hooves to the midpoint of the front hooves in the BEV coordinates. Let $\mathbf{d}$ denote the camera's forward direction projected onto the ground plane. The relative orientation angle θ between the cattle and the camera is computed as the difference between their angles with respect to the ground plane axes, calculated by the arctangent function:

$$\theta = \arctan\left(\frac{v_y}{v_x}\right) - \arctan\left(\frac{d_y}{d_x}\right), \quad (12)$$

where (v_x, v_y) and (d_x, d_y) are the components of the cattle orientation and camera direction vectors, respectively. This angle θ is discretized into four angular bins to assign a viewpoint ID, which is used to support viewpoint-aware retrieval. A visualization of this process is shown in Fig. 7.

5. Experiments

5.1. Dataset

We conduct experiments on the proposed dataset. To simulate practical deployment scenarios, we construct both open-set and closed-set ReID evaluation protocols. In the open-set setting, cattle identities appearing in the evaluation are completely disjoint from those in the training set, reflecting the need to recognize previously unseen individuals. In the closed-set setting, the same set of identities is shared between training and evaluation, allowing assessment under more controlled conditions.

Image resolution filtering. All images are cropped around individual cows based on annotated bounding boxes. To ensure sufficient visual information for ReID, we discard instance images whose width and height are both below 150 pixels. We visualize the resolution distribution of all cropped images in Fig. 8, where each point represents an image with its width and height. This filtering step removes severely low-resolution samples that may hinder feature learning.

Train-test partition and ID distribution. We construct both open-set and closed-set ReID evaluation protocols using data from non-overlapping cameras. In the closed-set setting, the same identities are used in both training and testing, enabling evaluation under controlled conditions. In contrast, the open-set setting uses disjoint identities between training and testing subsets, simulating realistic deployment scenarios where unseen individuals must be recognized.

For open-set evaluation, we construct multiple datasets with varying numbers of training identities, such as Imsil10 with 10 training IDs and Namwon9 with 9 training IDs. Note that each open-set subset uses only a portion of all the annotated images, and subsets may partially overlap.

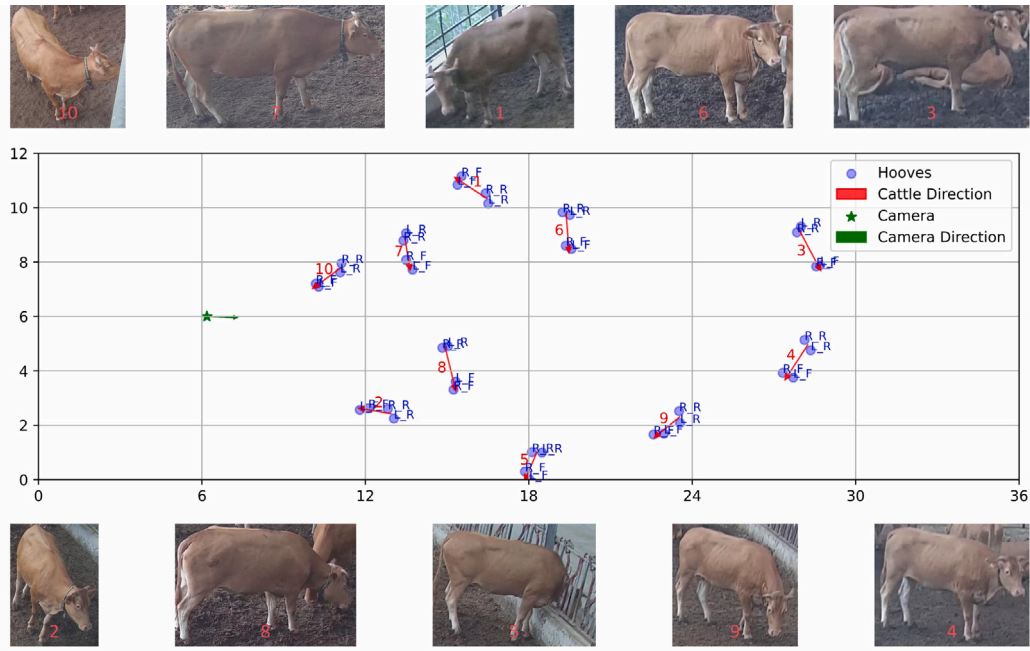


Fig. 7. Visualization of cattle orientation estimation from 2D pose keypoints. The central BEV plot illustrates projected hoof positions (blue), estimated orientation vectors (red), and camera direction (green). Surrounding images show the corresponding cattle instances for each vector, arranged according to their assigned numbers. The numerical labels in the BEV plot are not cattle IDs but are used solely for visual correspondence between each direction estimate and its associated instance image.

Table 3

Statistics of the constructed open-set and closed-set ReID datasets.

Type	Dataset	Train(IDs)	Train(Imgs)	Query(IDs)	Query(Imgs)	Gallery(IDs)	Gallery(Imgs)	Total(IDs)	Total(Imgs)
Open-set	Imsil10	10	1290	9	269	9	863	19	2422
	Imsil13	13	1689	6	237	6	739	19	2665
	Imsil16	16	2258	3	120	3	339	19	2717
	Namwon9	9	1989	3	96	3	406	12	2491
Closed-set	Imsil-day1	19	636	19	63	19	196	19	895
	Imsil-day2	19	1041	19	109	19	306	19	1456
	Imsil-day3	19	1444	19	152	19	419	19	2015
	Imsil-day4	19	1953	19	212	19	559	19	2724
	Namwon-cam4	3	611	3	67	3	172	3	850
	Namwon-cam5	9	1428	9	154	9	407	9	1989

To quantify the difficulty of each open-set configuration, we compute the test-to-total identity ratio (TTR), defined as:

$$TTR = \frac{N_{\text{test}}}{N_{\text{train}} + N_{\text{test}}} \quad (13)$$

This metric reflects the proportion of unseen identities during training, with higher TTR values indicating more challenging generalization scenarios. These controlled variations in TTR allow us to analyze the model's robustness under different levels of data scarcity.

Detailed statistics, including the number of identities and images in training, query, and gallery sets, are provided in Table 3.

5.2. Implementation details

We adopt a transformer-based architecture initialized from ImageNet-pretrained weights. Input images are resized to 256×256 to accommodate cattle appearance variations caused by pose and viewpoint differences.

To extract 2D pose features, we utilize a PHE consisting of two convolutional layers with kernel size 1. The first layer projects the single-channel input heatmap to an embedding dimension of 768, followed by a ReLU activation and a second convolutional layer maintaining the

same embedding dimension. This pose encoder thus produces a 768-dimensional feature representation for each spatial location in the pose heatmap.

To enhance model generalization, we apply data augmentation strategies including random horizontal flipping, random cropping with padding, and random erasing. All images are normalized using a mean and standard deviation of 0.5. The model is trained for up to 720 epochs, with a linear warmup strategy applied at the beginning of training. Model optimization is performed using stochastic gradient descent (SGD) with an initial learning rate of 0.001 and a weight decay of 1×10^{-4} . We use a batch size of 64 and optimize the network using a combination of softmax loss and triplet loss, with 8 instances per identity. All experiments are conducted on a single NVIDIA H100 GPU.

Our evaluation protocol is based on the standard Market-1501 (Zheng et al., 2015) framework, with two customized settings that reflect the characteristics of our dataset and deployment scenarios:

- **Cross-camera ReID:** This setting is used with the Imsil subset, which includes two cameras. Query and gallery images are sampled from different camera views to simulate real-world ReID scenarios across distributed surveillance systems. To ensure evaluation fairness, gallery images sharing the same identity and camera as the query are excluded.

Table 4

Comparison of methods under closed-set identification setting.

Backbone	Method	Imsil-day1		Imsil-day2		Imsil-day3		Imsil-day4		Namwon-cam4		Namwon-cam5	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ResNet-50	AGW	61.5	72.1	64.9	75.9	64.0	77.0	80.4	86.8	97.6	98.5	72.9	70.8
	TransReID	73.3	78.7	72.1	76.9	82.1	84.9	92.0	94.8	99.0	100.0	72.9	80.5
Hybrid	HAT	63.9	70.5	75.5	80.6	80.3	86.8	92.6	93.4	99.2	100.0	76.2	81.2
ViT-B/16	TransReID	86.6	88.5	76.4	78.7	84.6	88.2	93.3	93.4	97.9	98.5	78.2	83.8
	PFD	83.3	88.5	76.8	80.6	81.5	87.5	93.2	93.9	99.5	100.0	78.6	83.1
	RotTrans	84.3	86.9	77.2	80.6	84.8	88.8	94.6	94.8	96.0	98.5	76.2	82.5
	PHA	83.2	85.2	78.4	80.6	84.4	87.5	92.3	94.3	99.6	98.5	62.8	77.3
	Ours (PHE only)	88.3	91.8	79.0	81.5	87.9	89.5	94.0	95.3	99.9	100.0	80.6	84.4
	Ours (PHE + VCR)	94.0	92.7	81.5	79.1	89.4	87.5	95.3	96.8	99.9	100.0	83.1	86.9

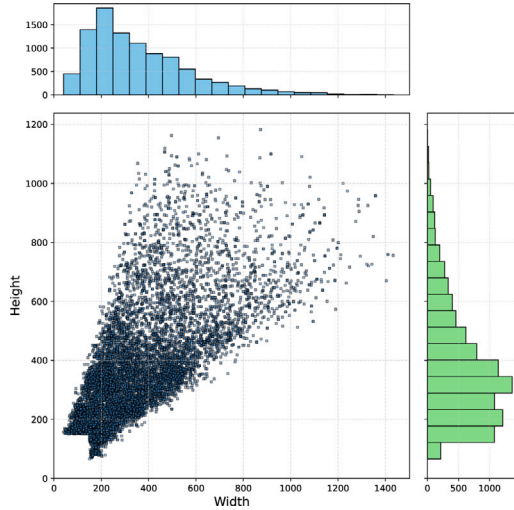


Fig. 8. Distribution of cropped image resolutions for individual cow instances in the dataset. Each blue dot represents a cropped image plotted by width along the x-axis and height along the y-axis. The marginal histograms above and to the right show the frequency distribution of widths and heights, respectively. Images with extremely low resolutions below the filtering threshold are removed to ensure sufficient visual quality for ReID.

- Same-camera ReID: This setting is applied to the Namwon subset, which contains only one single camera. Query and gallery images are drawn from the same camera view, allowing us to evaluate the model's within-view feature discrimination capabilities. In this case, gallery images from the same identity but different time points are used, and cross-camera samples are not applicable.

We report standard ReID evaluation metrics, including mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-1 and Rank-5, to reflect retrieval accuracy under varying degrees of challenge. The two evaluation settings, cross-camera and same-camera, also reflect realistic deployment scenarios such as distributed farm surveillance or fixed-camera local monitoring.

5.3. Performance evaluation

Closed-set evaluation. We begin by evaluating our method under the closed-set identification setting. Table 4 compares our models against several representative baselines across six subsets: four from Imsil (day1–day4) and two from Namwon (cam4 and cam5). For the Imsil subsets, we adopt cross-camera evaluation to better reflect practical deployment across multiple camera viewpoints. In contrast, for the Namwon subsets, we employ same-camera evaluation, leveraging the consistent viewpoint to model realistic scenarios in fixed-camera surveillance. The compared methods include CNN-based approaches

(AGW (Ye et al., 2021), TransReID (He et al., 2021) with ResNet-50 (He et al., 2016)), a hybrid architecture (HAT (Zhang et al., 2021)), and transformer-based models (TransReID (He et al., 2021), PFD (Wang et al., 2022), RotTrans (Chen et al., 2022), PHA (Zhang et al., 2023) with ViT-B/16 (Dosovitskiy et al., 2020) backbone).

Our approaches, Ours (PHE only) and Ours (PHE + VCR), consistently outperform all baselines in both mAP and Rank-1 accuracy, demonstrating strong performance across various conditions. The Imsil subsets pose a greater challenge due to the presence of 19 individual cattle, requiring fine-grained discrimination. Notably, Ours (PHE + VCR) achieves the highest scores, e.g., 94.0% mAP and 92.7% Rank-1 on Imsil-day1. On the largest closed-set subset, Imsil-day4, which contains 2724 images, our method achieves up to 95.3% mAP and 96.8% Rank-1, indicating the effectiveness of incorporating the VCR module in complex identification scenarios.

Due to the small number of identities in Namwon-cam4, performance differences among methods are marginal, and most achieve saturated results, including ours. Nevertheless, the Namwon environment introduces practical challenges, such as mild occlusions caused by structural elements in the cattle farm. Specifically, iron beams positioned above the fences can partially obscure the animals, as illustrated in Fig. 3. Despite these visual distractions, our models maintain robust identification accuracy, highlighting their applicability in real-world deployments.

Open-set evaluation. We further evaluate the models under the open-set setting to assess their ability to generalize to unseen cattle identities. As the number of available cameras varies across different subsets, we adopt a same-camera ReID protocol for all open-set evaluations to ensure consistency and fair comparison. Moreover, benefiting from the inherent multi-view nature of our dataset, where each Hanwoo cattle appears in diverse poses and orientations, the same-camera setting still provides sufficient appearance variation to effectively evaluate model generalization. Based on the statistics summarized in Table 3, we construct four open-set subsets with varying TTR levels. As shown in Table 5, we compare our method with several ViT-based baselines that perform well in the closed-set setting.

Compared to the baseline methods, our approaches demonstrate generally superior performance across various open-set configurations. On Imsil10, which has a relatively high TTR of 0.47, the open-set task becomes more challenging due to the limited number of training identities. Nevertheless, Ours (PHE only) achieves the highest Rank-5 accuracy of 97.0% and a strong mAP of 54.9%, outperforming TransReID and RotTrans. The enhanced Ours (PHE + VCR) further boosts mAP to 60.1% while maintaining the best Rank-1 accuracy (85.7%), demonstrating the added benefit of incorporating the VCR module.

As the number of training identities increases from Imsil10 to Imsil16, the training set provides the model with richer identity supervision. Under this setting, our method consistently achieves the highest mAP scores across all Imsil subsets, with the PHE + VCR combination outperforming all other methods. In particular, on Imsil16 (TTR = 0.16), Ours (PHE + VCR) attains the highest mAP of 74.8% and a competitive Rank-1 accuracy of 86.7%, demonstrating the effectiveness of

Table 5

Comparison of methods under the open-set identification setting. All methods are implemented using a ViT-B/16 backbone.

Method	Imsil10 (TTR = 0.47)			Imsil13 (TTR = 0.32)			Imsil16 (TTR = 0.16)			Namwon9 (TTR = 0.25)		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
TransReID	53.3	84.8	93.7	60.6	81.9	95.4	70.7	85.8	96.7	62.8	82.3	90.6
RotTrans	54.3	85.1	95.5	60.2	83.5	94.9	71.0	86.7	96.7	63.6	83.3	93.8
PHA	43.1	80.3	92.9	59.0	82.3	93.2	67.4	86.7	94.2	63.3	83.3	92.7
Ours (PHE only)	54.9	85.1	97.0	61.5	85.2	96.2	71.4	87.5	97.5	64.5	89.6	95.8
Ours (PHE + VCR)	60.1	85.7	96.1	62.8	81.0	93.7	74.8	86.7	95.0	68.7	86.5	94.8

our pose-guided design combined with VCR in capturing discriminative cues. For Namwon9, a cross-scene evaluation involving training and testing on different cameras, Ours (PHE + VCR) achieves the best mAP (68.7%) among all methods, while Ours (PHE only) attains the highest Rank-1 accuracy (89.6%). These results highlight the robustness of our framework to both intra-scene and inter-scene domain shifts. Although some rank-based metrics are comparable across methods, our approach consistently balances identification accuracy and generalization ability through the incorporation of pose-aware features.

It is worth noting that while mAP consistently improves after introducing the VCR module, rank-based metrics (e.g., Rank-1, Rank-5) exhibit mixed trends across different datasets. This discrepancy stems from the fundamental differences in how these metrics evaluate retrieval performance: mAP measures the average ranking of all true positive matches, rewarding models that push more correct matches toward the top of the list, whereas Rank-1 considers only whether the very first retrieved sample is correct.

The occasional decrease in Rank-1 accuracy after incorporating the VCR module can be attributed to its filtering strategy. During evaluation, VCR removes gallery images with inconsistent viewpoints relative to the query, discarding cross-view positive matches that differ substantially in appearance due to view discrepancy. These matches are sometimes ranked highly not because the model accurately captures the cattle's appearance, but because it inadvertently relies on background elements such as ground textures or scenery (e.g., the fourth row in Fig. 10). By eliminating these misleading matches, VCR encourages the model to focus more precisely on discriminative features of the cattle itself. As a result, the remaining true positives tend to be semantically better aligned with the query, leading to improved mAP. However, since Rank-1 evaluates only the top-ranked match, it can decrease if the removed cross-view positive was previously occupying the first position and is not replaced by another correct match. This trade-off highlights how our method enhances overall retrieval quality, even if the immediate top match occasionally changes.

Training dynamics. To gain deeper insights into model behavior during training, we analyze the mAP-loss trends of different algorithms, as shown in Fig. 9. Each curve represents the evolution of mAP as training loss decreases. A semi-transparent circle marks the peak mAP achieved by each method, with its area proportional to the corresponding Rank-1 accuracy, enabling a joint comparison of retrieval performance and convergence dynamics. Note that in all subfigures, the label Ours refers to our full model with both PHE and VCR enabled.

On the Imsil-day1 subset, our method shows a consistently upward trend in mAP as the loss decreases, ultimately achieving the highest mAP and Rank-1 accuracy among all methods. This indicates both efficient optimization and strong generalization. In contrast, methods such as AGW and HAT exhibit unstable mAP trajectories, with noticeable fluctuations even at later training stages, suggesting limited robustness and convergence issues under the challenging fine-grained setting posed by Imsil.

On the Namwon-cam5 subset, which features fewer identities but a more visually cluttered background, the performance gap between methods narrows slightly. Nevertheless, our method remains the best overall performer, reaching the highest mAP and Rank-1. Notably, some methods (e.g., PFD and HAT) peak at relatively high loss values,

which could reflect early overfitting or sensitivity to specific visual patterns. In contrast, our model steadily improves as training progresses, showing both stability and efficiency.

On the open-set subsets shown in Fig. 9(c) and Fig. 9(d), similar trends are observed. In Fig. 9(c), our method again outperforms all baselines on Imsil13, with a smooth mAP increase throughout training and a peak marked by the largest circle, indicating superior Rank-1 accuracy. Notably, RotTrans exhibits severe instability, with significant mAP drops during training, suggesting poor generalization in open-set conditions. TransReID and PHA show moderate performance, but with less consistent progression.

As shown in Fig. 9(d), which presents results on the Namwon9 dataset, our method consistently outperforms other methods throughout the training process, achieving a final mAP of approximately 68%. In contrast, both TransReID and PHA exhibit relatively stable but limited improvements, eventually plateauing at around 62% and 63%, respectively. Notably, RotTrans suffers from severe instability, with frequent and sharp drops in mAP as training progresses. These fluctuations highlight its difficulty in achieving robust convergence. Overall, our method demonstrates superior robustness and generalization, maintaining high performance even as the loss decreases, and avoids performance degradation in the later stages of training.

Qualitative visualization. To further illustrate the effectiveness of our method, we compare the attention maps generated by different models in Fig. 10. The visualization reveals that our model consistently focuses on semantically meaningful regions. In contrast, baseline methods such as PHA and TransReID often attend to overly broad or irrelevant areas, including background clutter or the ground.

This focused attention can be attributed to the integration of pose keypoints, which guide the model toward anatomically informative regions. As highlighted by the yellow and purple dashed ellipses, our attention maps are not only more compact but also better aligned with the target individual. This results in enhanced spatial precision and improved robustness to background interference.

To better understand the model's retrieval behavior, we visualize the top-5 results for selected query identities in Fig. 11. Each row presents a query image (leftmost) followed by its top-5 ranked gallery images. Correct matches are framed in green, while incorrect ones are in red.

All visualized samples are drawn from gallery images with the same ViewID as the query, where the ViewID denotes the estimated relative orientation of the individual with respect to the camera. This alignment mitigates variations caused by drastic viewpoint changes, allowing us to focus on intra-view retrieval performance. Despite the shared viewpoints, the top-5 candidates often appear highly similar, making fine-grained recognition challenging.

We observe that even within the same viewpoint, cattle can exhibit significant appearance variation due to pose and occlusion. For example, when a cow faces away from the camera, only the hindquarters are visible, while a frontal view reveals the head and facial details. Some gallery images also suffer from occlusion, such as cows feeding through fences, which can obscure the head and lead to visually misleading matches (e.g., ID17 in Row 5, Rank-4).

Most retrievals correctly rank the same identity at high positions. However, errors still occur when different individuals share highly

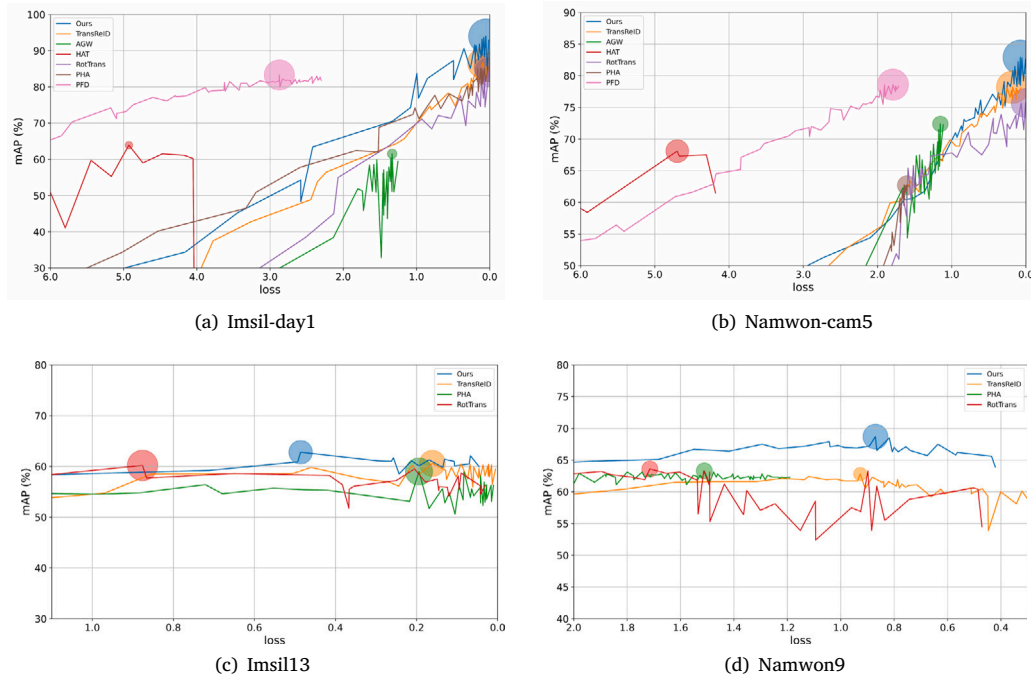


Fig. 9. Comparison of performance trends in terms of mAP and training loss across closed-set and open-set settings. Fig. (a) and Fig. (b) show results on closed-set datasets (Imsil-day1 and Namwon-cam5), while Fig. (c) and Fig. (d) show results on open-set datasets (Imsil13 and Namwon9). Each curve illustrates the progression of mAP against decreasing training loss for different methods. A translucent circle marks the peak mAP of each method, with the size of the circle proportional to the corresponding Rank-1 accuracy, enabling a joint analysis of retrieval performance and convergence behavior.

similar visual traits. For instance, ID16 (Row 4) has no horns, while its Rank-4 and Rank-5 mismatches both have visible horns. Conversely, ID14 (Row 2) has horns, but its Rank-5 match lacks them. These examples highlight the challenges posed by fine-grained visual differences and partial occlusions in cattle ReID.

5.4. Ablation study

We conduct ablation studies to analyze the effectiveness of key components in our proposed framework. Note that the baseline model used throughout our comparisons is TransReID (He et al., 2021), a strong ViT-based ReID method. In Section 5.3, we have already compared our full model against this baseline in terms of mAP and Rank-1 accuracy across various closed-set and open-set settings.

In particular, TransReID serves as our baseline without any pose-related cues or explicit viewpoint filtering. While the original TransReID framework includes an optional Viewpoint Embedding module, which assigns different learnable embeddings based on predefined view labels, we found in our internal exploration that directly incorporating such ViewID embeddings did not lead to consistent improvements in the cattle ReID scenario. As a result, this component was not included in the main comparisons presented earlier in the paper.

To provide a fair and comprehensive evaluation, we further compare our proposed ViewID-based filtering strategy with the original Viewpoint Embedding mechanism from TransReID in the following ablation study. This allows us to assess the relative effectiveness of our approach in leveraging viewpoint information for improved identification performance.

Comparative analysis of viewpoint encoding methods. Table 6 presents an ablation study on three components for modeling viewpoint variations: the pose heatmap encoder (PHE), viewID embedding (VID), and viewpoint-constrained retrieval (VCR). Among them, PHE encodes pose-dependent visual cues via keypoint heatmaps, while VCR restricts retrieval to gallery images with consistent viewpoint. In contrast, VID,

a module introduced in TransReID (He et al., 2021), is included in our baseline and models viewpoint information via a learnable embedding vector.

While the combination of PHE and VCR achieves competitive performance, it consistently yields the best mAP across all datasets. For instance, on Imsil10, this combination achieves the highest mAP of 60.1%, outperforming the baseline (53.3%) by +6.8%. On Namwon9, it further improves the baseline from 62.8% to 68.7%, indicating its effectiveness in both intra-scene and cross-scene retrieval tasks. Notably, on Namwon9, using only PHE results in the highest Rank-1 (89.6%) and Rank-5 (95.8%) accuracies, demonstrating the strong discriminative power of pose-guided features.

The VID module, although designed to capture viewpoint variations, does not provide consistent gains in our cattle ReID setting. In fact, incorporating VID into the baseline often leads to performance drops, such as a decrease in mAP from 53.3% to 51.8% on Imsil10 and from 62.8% to 59.7% on Namwon9. This suggests that relying solely on learnable ViewID embeddings may interfere with the fine-grained spatial cues from pose heatmaps, and is insufficient for modeling complex orientation variations in animals.

VCR proves particularly effective for both recognition accuracy and practical deployment. In real-world applications such as precision livestock farming, VCR helps filter out gallery candidates with incompatible viewpoints, reducing false matches caused by large appearance differences. This leads to improved system robustness and enhanced user confidence. Moreover, VCR highlights the importance of viewpoint consistency in cross-camera ReID and serves as a lightweight and scalable solution for real-time applications.

The practical value of VCR is further supported by results in Table 4. In closed-set scenarios, especially in controlled environments such as indoor farms, VCR-based configurations consistently boost both mAP and rank-based metrics, reaffirming its utility in enhancing recognition reliability.

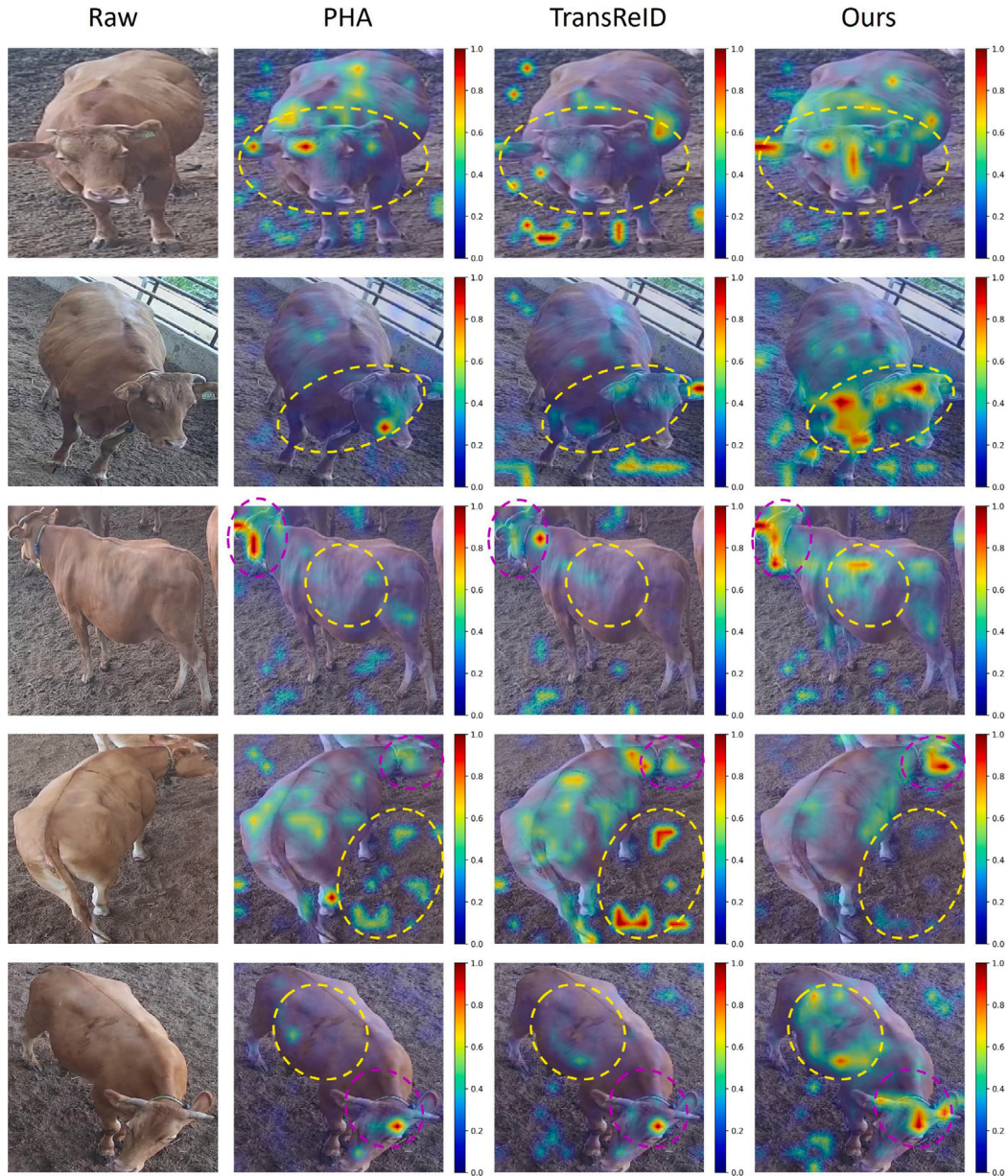


Fig. 10. Qualitative comparison of attention maps generated by different methods, including the raw image, PHA, TransReID, and our method. Notice that important regions are marked with yellow and purple dashed ellipses. Benefiting from pose keypoints, our method generates attention maps that are more concentrated on discriminative textures of the cow, while suppressing irrelevant background regions such as the ground or surrounding environment.

Table 6

Performance comparison of different viewpoint modeling module combinations. Columns indicate whether each component is used: Pose Heatmap Encoder (PHE), ViewID Embedding (VID), and Viewpoint-Constrained Retrieval (VCR).

PHE	VID	VCR	Imsil10 (TTR = 0.47)			Imsil13 (TTR = 0.32)			Namwon9 (TTR = 0.25)		
			mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
			53.3	84.8	93.7	60.6	81.9	95.4	62.8	82.3	90.6
	✓		51.8	85.1	94.1	60.2	82.3	93.7	59.7	84.4	92.7
	✓	✓	56.5	81.0	94.1	61.1	78.5	91.6	63.9	86.5	89.6
		✓	58.1	80.3	93.7	62.4	80.2	92.4	67.0	84.4	91.7
✓	✓		52.3	85.5	94.4	58.6	82.3	92.0	63.7	86.5	89.6
✓			54.9	85.1	97.0	61.5	85.2	96.2	64.5	89.6	95.8
✓	✓	✓	56.8	80.3	91.4	61.0	78.9	92.8	67.3	80.2	90.6
✓		✓	60.1	85.7	96.1	62.8	81.0	93.7	68.7	86.5	94.8



Fig. 11. Top-5 retrieval results for selected query identities. Each row shows a query image followed by its top-5 ranked gallery images. Correct matches are marked with green borders, and incorrect ones with red. This illustrates the model's retrieval accuracy and typical failure cases.

Table 7

Comparison of aggregation strategies for pose heatmaps.

Method	Imsil-day1 (Closed-set)			Imsil10 (TTR = 0.47)			Namwon9 (TTR = 0.25)		
	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
Max	94.0	92.7	97.6	60.1	85.7	96.1	68.7	86.5	94.8
Mean	91.7	87.8	95.1	59.9	80.3	94.1	64.6	81.2	92.7
Concat	91.7	90.2	95.1	43.3	64.3	85.1	62.3	80.2	91.7

Aggregation strategies for pose heatmap fusion. Table 7 compares three different strategies for aggregating pose heatmaps as input to the Pose Heatmap Encoder (PHE): max pooling, average pooling (mean), and concatenation. These strategies determine how the multiple keypoint heatmap channels are combined before being encoded.

The results show that max pooling consistently outperforms other strategies across all datasets. It achieves the best performance on Imsil-day1 (mAP 94.0%, Rank-1 92.7%) and Namwon9 (mAP 68.7%, Rank-1 86.5%), as well as the highest scores on Imsil10 (mAP 60.1%, Rank-1 85.7%), demonstrating its effectiveness in both closed-set and open-set scenarios.

Mean pooling performs comparably but slightly worse, especially on Namwon9 (mAP 64.6%, Rank-1 81.2%), suggesting that averaging spatial responses may dilute discriminative pose information. Concatenation, although it preserves all channel information by stacking them together, consistently yields the worst performance. For example, on Imsil10, it results in a significantly lower mAP of 43.3% and Rank-1 accuracy of 64.3%. This indicates that increasing dimensionality without aggregation may introduce optimization challenges or lead to overfitting.

In summary, this ablation validates that max pooling is the most effective and robust aggregation strategy for pose heatmaps in our

setting. While mean pooling may serve as a lightweight alternative, concatenation is not recommended due to its suboptimal performance and computational inefficiency.

Robustness to pose keypoint errors. In this section, we evaluate the robustness of our framework to errors in pose keypoint detection, which may arise due to inherent limitations of the pose estimation network, particularly in the presence of occlusion or complex postures.

To assess the reliability of the pose estimator, we first visualize keypoints extracted under various real-world conditions. As shown in Fig. 12, the network consistently detects most keypoints, even when cattle are lying down, feeding, or partially occluded. These qualitative results highlight the estimator's generalizability and its practical applicability in unconstrained environments.

To accommodate incomplete or missing keypoints, PHE operates on dense Gaussian heatmaps rather than sparse keypoint coordinates. Each detected keypoint is encoded as a separate heatmap channel, with missing keypoints corresponding to zero-valued channels. These are aggregated into a single heatmap, which is fused with image features via element-wise addition. As a result, missing keypoints contribute no information, while inaccurately localized ones may introduce noise. Nevertheless, the spatial structure of the heatmaps provides a degree

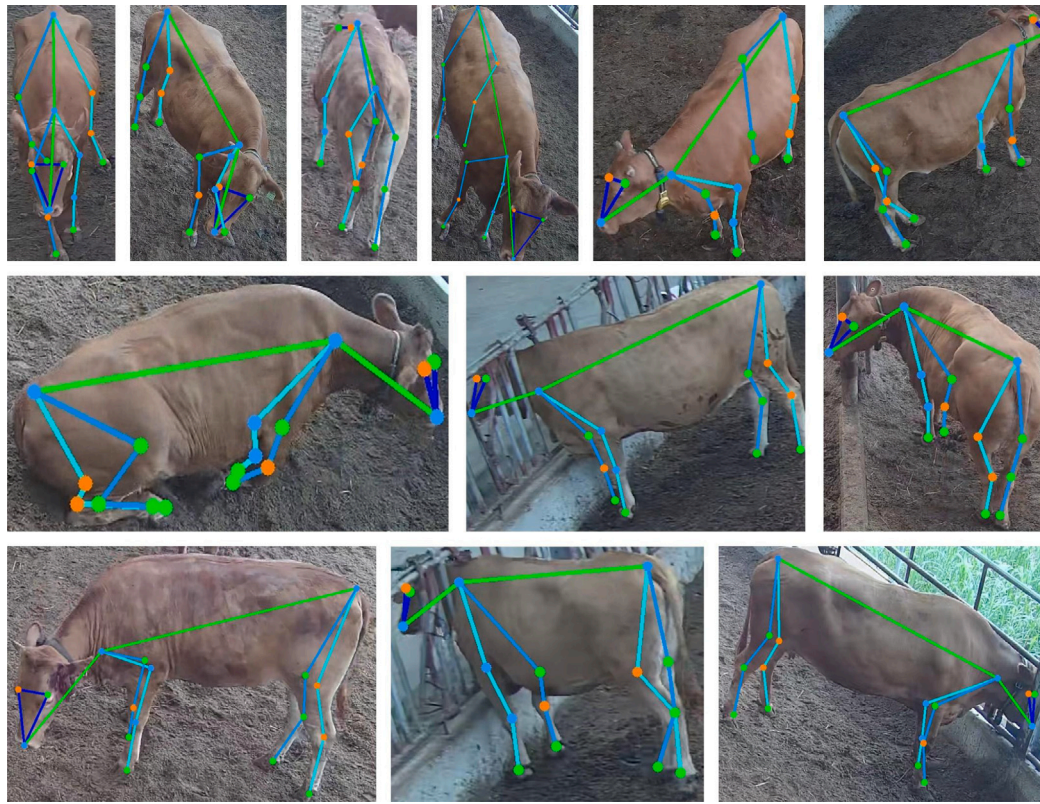


Fig. 12. Examples of pose keypoint detection results under various cattle postures, including lying down, feeding, and partially occluded cases.

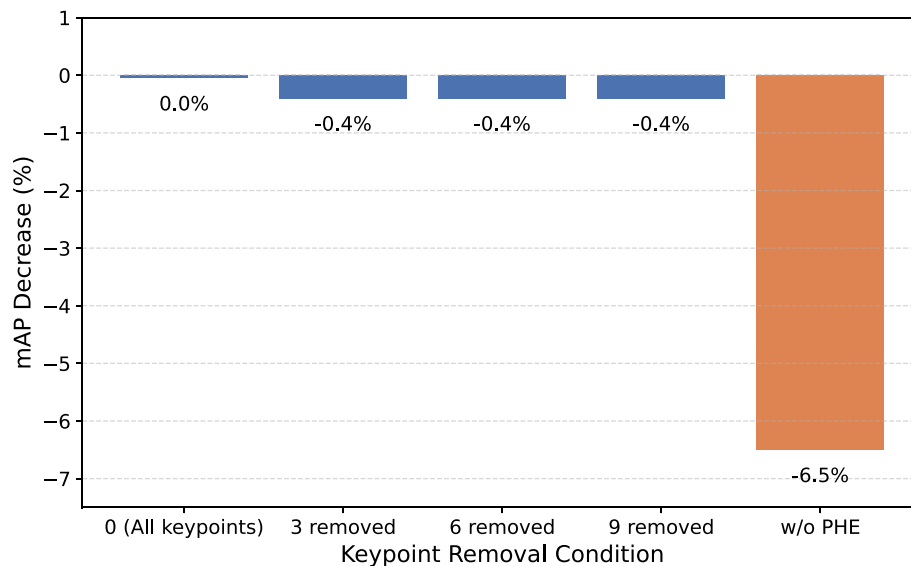


Fig. 13. Impact of random keypoint removal on the mAP performance of the PHE module on the Imsil-day1 dataset. The labels “0”, “3 removed”, “6 removed”, “9 removed”, and “w/o PHE” correspond to the number of keypoints removed and the condition without the PHE module, respectively.

of tolerance to such errors, enhancing robustness without over-reliance on pose inputs.

We further evaluate the robustness of PHE through an ablation study, in which varying numbers of keypoints are randomly removed before being fed into the module. As presented in Fig. 13, PHE maintains stable performance even when approximately 53% of the keypoints (9 out of 17) are removed, showing only a moderate drop in mAP. In contrast, completely disabling the PHE module leads to a

significant performance decline of 6.5%, highlighting its overall contribution to the framework. These results confirm the module’s resilience to partial input degradation.

For VCR, robustness has already been validated in our earlier ablation study, with results summarized in Table 6. These experiments inherently account for real-world pose estimation inaccuracies, such as missing or imprecise keypoints. Despite such imperfections, VCR consistently improves ReID performance over the baseline. In rare

Table 8
Inference efficiency under different batch sizes with PHE enabled.

Batch size	Time per Image (s)	FPS	Notes
1	0.14347	69.70	No batching
32	0.00770	129.93	GPU-parallelized inference
64	0.00748	133.71	Slight gain from larger batch

cases of detection failure (e.g., any missing hoof keypoints), VCR skips viewpoint estimation for the affected image. Otherwise, it remains effective even in the presence of minor keypoint localization errors.

Overall, these mechanisms ensure that our framework remains robust even in the presence of partial or complete pose estimation errors, allowing stable ReID performance without interruption. While deviations in keypoint detection may introduce some noise, our method is designed to maximize robustness and error tolerance, mitigating the impact of such inaccuracies as much as possible.

Inference efficiency and real-time deployment capability. Our proposed framework is designed in a modular fashion, enabling the ReID component to be flexibly integrated into comprehensive multi-task systems that may include object detection, tracking, and action recognition. This modular design allows the ReID functionality to be selectively applied without introducing significant computational overhead to the overall system.

To assess inference-time efficiency, we separately measure the computational cost of the PHE and the VCR modules. As summarized in Table 8, the inference speed varies with the batch size. Under batch size 1, which simulates single-frame inference or edge-device deployment, the system processes each image in 0.14347 s, corresponding to 69.70 frames per second (FPS). When the batch size is increased to 32 or 64, which is more representative of server-based or batch processing scenarios, the FPS rises to 129.93 and 133.71, respectively. These results exceed the commonly accepted threshold of 60 FPS for real-time performance.

The VCR module is extremely lightweight, requiring only 0.00013 s per individual to compute the ViewID, resulting in an effective processing speed exceeding 7000 FPS. The computational cost of this module is therefore negligible in the context of the overall inference pipeline.

These findings demonstrate that while the PHE module introduces moderate computational overhead, the framework as a whole remains capable of real-time performance across various deployment environments. The proposed system is thus well-suited for practical applications requiring rapid and modular cattle identification in multi-camera and multi-view settings.

6. Conclusion

This paper presents a deployment-friendly framework for Hanwoo cattle ReID in real-world farm environments. Unlike methods that rely on extracting specific views such as facial or muzzle images, our framework is pose-free and requires only the installation of cameras in barns without manual viewpoint selection, enabling practical and scalable deployment in precision livestock farming scenarios. To address challenging pose and viewpoint variations, we introduce two novel modules: the PHE, which leverages 2D keypoints from a pose estimation model to guide the Transformer’s attention toward identity-relevant regions, improving robustness to body deformation; and the VCR strategy, which estimates cattle orientation from monocular images by projecting hoof keypoints onto a BEV plane and filtering gallery samples based on viewpoint consistency, thereby enhancing retrieval accuracy under large orientation changes.

We also contribute the HanwooReID dataset, which contains 9,929 manually verified images of 31 Hanwoo cattle. Collected in authentic barn environments, the dataset captures natural variations in cattle poses and viewpoints, offering a representative and practical resource

for developing and evaluating cattle ReID methods under real-world farm conditions. Extensive experiments under both closed-set and open-set evaluation settings demonstrate the effectiveness of our proposed framework. In closed-set scenarios, our method achieves up to 7.4% improvement in mean Average Precision (mAP) over our baseline. In open-set settings, where the test identities are entirely disjoint from the training set, we observe up to 6.8% mAP gain, highlighting the framework’s generalizability to unseen individuals.

Despite these promising results, certain limitations remain. While infrared images offer clear contour visibility in low-light or nighttime environments, their lack of color and limited texture contrast make individual discrimination more challenging, especially for Hanwoo cattle whose external appearance is relatively homogeneous. Consequently, our current work focuses on daytime RGB images, which preserve richer visual cues such as coat color, texture, and shadows. In future work, we plan to incorporate IR-based data and explore cross-temporal ReID to improve robustness under varying illumination conditions. We also aim to expand the dataset with more identities, larger image collections, and short video sequences to capture subtle behavioral patterns, which may serve as a complementary cue for identity discrimination. By addressing these directions, we believe our framework can further advance the applicability of cattle ReID in precision livestock farming.

CRediT authorship contribution statement

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

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT by OpenAI in order to improve the clarity and fluency of the English language. After using this tool, the author(s) reviewed and edited the content as needed and take full responsibility for the content of the publication.

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 National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT)(RS-2024-00392406) and (2020R1A2C2013060); by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2019-NR040079); and by the computing resources and infrastructure provided by the Micro Data Center at Mokpo National University in 2025.

Data availability

Data will be made available on request.

References

- Ahmed, S.U., Frnda, J., Waqas, M., Khan, M.H., 2024. Dataset of cattle biometrics through muzzle images. *Data Brief* 53, 110125.
- Anon, 2025. CVAT: Computer vision annotation tool. <https://github.com/cvat-ai/cvat>. (Accessed 29 July 2025).
- Awad, A.I., 2016. From classical methods to animal biometrics: A review on cattle identification and tracking. *Comput. Electron. Agric.* 123, 423–435. <http://dx.doi.org/10.1016/j.compag.2016.03.014>.
- Bakhshayeshi, I., Erfani, E., Taghikhah, F.R., Elbourn, S., Beheshti, A., Asadnia, M., 2023. An intelligence cattle reidentification system over transport by siamese neural networks and yolo. *IEEE Internet Things J.* 11 (2), 2351–2363.
- Bergamini, L., Porrello, A., Dondona, A.C., Del Negro, E., Mattioli, M., D'alterio, N., Calderara, S., 2018. Multi-views embedding for cattle re-identification. In: 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems. SITIS, IEEE, pp. 184–191.
- Bergman, N., Yitzhaky, Y., Halachmi, I., 2024. Biometric identification of dairy cows via real-time facial recognition. *Animal* 18 (3), 101079.
- Bruslund Haurum, J., Karpova, A., Pedersen, M., Hein Bengtson, S., Moeslund, T.B., 2020. Re-identification of zebrafish using metric learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. pp. 1–11.
- Cao, J., Tang, H., Fang, H.-S., Shen, X., Lu, C., Tai, Y.-W., 2019. Cross-domain adaptation for animal pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9498–9507.
- Chang, A.Z., Fogarty, E.S., Moraes, L.E., García-Guerra, A., Swain, D.L., Trotter, M.G., 2022. Detection of rumination in cattle using an accelerometer ear-tag: A comparison of analytical methods and individual animal and generic models. *Comput. Electron. Agric.* 192, 106595.
- Chen, S., Ye, M., Du, B., 2022. Rotation invariant transformer for recognizing object in UAVs. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 2565–2574.
- Coffman, E., Clark, R., Bui, N.-T., Pham, T.T., Kegley, B., Powell, J.G., Zhao, J., Le, N., 2024. Cattleface-rgb-t: Rgb-t cattle facial landmark benchmark. *arXiv preprint arXiv:2406.03431*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dulal, R., Zheng, L., Kabir, M.A., 2025. MHAFF: Multi-head attention feature fusion of CNN and transformer for cattle identification. *arXiv preprint arXiv:2501.05209*.
- Fuentes, A., Yoon, S., Park, J., Park, D.S., 2020. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. *Comput. Electron. Agric.* 177, 105627.
- Gao, J., Burghardt, T., Andrew, W., Dowsey, A.W., Campbell, N.W., 2021. Towards self-supervision for video identification of individual holstein-friesian cattle: The cows2021 dataset. *arXiv preprint arXiv:2105.01938*.
- He, S., Luo, H., Wang, P., Wang, F., Li, H., Jiang, W., 2021. Transreid: Transformer-based object re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15013–15022.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.
- Jiang, T., Lu, P., Zhang, K., Ma, N., Han, R., Lyu, C., Li, Y., Chen, K., 2023. RtmPose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*.
- Junping, Y., Zebin, Y., Xiaojun, L., Yi, G., 2023. Pose-guided self and external attention feature matching and aggregation network for person re-identification. *Displays* 80, 102567.
- Kishore, R., Aslam, N., Kolekar, M.H., 2024. Patreid: Pose apprise transformer network for vehicle re-identification. *IEEE Trans. Emerg. Top. Comput. Intell.*
- Kumar, S., Singh, S.K., 2020. Cattle recognition: A new frontier in visual animal biometrics research. *Proc. Natl. Acad. Sci. India Sect. A: Phys. Sci.* 90 (4), 689–708.
- Li, S., Fu, L., Sun, Y., Mu, Y., Chen, L., Li, J., Gong, H., 2021. Individual dairy cow identification based on lightweight convolutional neural network. *PLoS One* 16 (11), e0260510.
- Li, Z., Ge, C., Shen, S., Li, X., 2018. Cow individual identification based on convolutional neural network. In: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence. pp. 1–5.
- Li, J., Zou, X., Wang, S., Chen, B., Xing, J., Tao, P., 2024. A parallel attention network for cattle face recognition. *arXiv preprint arXiv:2403.19980*.
- Lin, Y., Liu, L., Shi, J., 2024. Categorical keypoint positional embedding for robust animal re-identification. *arXiv preprint arXiv:2412.00818*.
- Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S., 2017. End-to-end comparative attention networks for person re-identification. *IEEE Trans. Image Process.* 26 (7), 3492–3506.
- Liu, J., Ni, B., Yan, Y., Zhou, P., Cheng, S., Hu, J., 2018. Pose transferrable person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4099–4108.
- Liu, Y., Wang, T., Zhang, X., Sun, J., 2022. Petr: Position embedding transformation for multi-view 3d object detection. In: European Conference on Computer Vision. Springer, pp. 531–548.
- Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X., 2023. PetrV2: A unified framework for 3d perception from multi-camera images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3262–3272.
- Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., 2019. Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops.
- Moskvyak, O., Maire, F., Dayoub, F., Armstrong, A.O., Baktashmotlagh, M., 2021. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings. In: 2021 Digital Image Computing: Techniques and Applications. DICTA, IEEE, pp. 1–8.
- Nasir, M.F., Fuentes, A., Han, S., Liu, J., Jeong, Y., Yoon, S., Park, D.S., 2025. Multi-camera fusion and bird-eye view location mapping for deep learning-based cattle behavior monitoring. *Artif. Intell. Agric.*
- Ong, K.E., Retta, S., Srinivasan, R., Tan, S., Liu, J., 2023. Cattleeyeview: A multi-task top-down view cattle dataset for smarter precision livestock farming. In: 2023 IEEE International Conference on Visual Communications and Image Processing. VCIP, IEEE, pp. 1–5.
- Pretto, A., Savio, G., Gottardo, F., Ucheddu, F., Concheri, G., 2024. A novel low-cost visual ear tag based identification system for precision beef cattle livestock farming. *Inf. Process. Agric.* 11 (1), 117–126.
- Simanungkalit, G., Barwick, J., Cowley, F., Dawson, B., Dobos, R., Hegarty, R., 2021. Use of an ear-tag accelerometer and a radio-frequency identification (RFID) system for monitoring the licking behaviour in grazing cattle. *Appl. Anim. Behav. Sci.* 244, 105491.
- Wang, L., Ding, R., Zhai, Y., Zhang, Q., Tang, W., Zheng, N., Hua, G., 2021. Giant panda identification. *IEEE Trans. Image Process.* 30, 2837–2849.
- Wang, T., Liu, H., Song, P., Guo, T., Shi, W., 2022. Pose-guided feature disentangling for occluded person re-identification based on transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2540–2549.
- Weng, Z., Meng, F., Liu, S., Zhang, Y., Zheng, Z., Gong, C., 2022. Cattle face recognition based on a two-branch convolutional neural network. *Comput. Electron. Agric.* 196, 106871.
- Xu, X., Deng, H., Shang, Y., Wang, Y., Zhang, S., Chen, D., Song, H., 2025. Multi-camera multi-cow tracking under non-overhead views. *Expert Syst. Appl.* 291, 128519. <http://dx.doi.org/10.1016/j.eswa.2025.128519>, URL <https://www.sciencedirect.com/science/article/pii/S0957417425021384>.
- Xu, Y., Zhang, J., Zhang, Q., Tao, D., 2022. Vitpose: Simple vision transformer baselines for human pose estimation. *Adv. Neural Inf. Process. Syst.* 35, 38571–38584.
- Ye, M., Chen, S., Li, C., Zheng, W.-S., Crandall, D., Du, B., 2024. Transformer for object re-identification: A survey. *Int. J. Comput. Vis.* 1–31.
- Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H., 2021. Deep learning for person re-identification: A survey and outlook. *IEEE Trans. Pattern Anal. Mach. Intell.*
- Yu, P., Burghardt, T., Dowsey, A.W., Campbell, N.W., 2024. MultiCamCows2024—a multi-view image dataset for AI-driven holstein-friesian cattle re-identification on a working farm. *arXiv preprint arXiv:2410.12695*.
- Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D., 2021. Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*.
- Zhang, G., Zhang, P., Qi, J., Lu, H., 2021. Hat: Hierarchical aggregation transformers for person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 516–525.
- Zhang, G., Zhang, Y., Zhang, T., Li, B., Pu, S., 2023. PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 14133–14142.
- Zheng, L., Huang, Y., Lu, H., Yang, Y., 2019. Pose-invariant embedding for deep person re-identification. *IEEE Trans. Image Process.* 28 (9), 4500–4509.
- Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1116–1124.
- Zheng, Z., Zheng, L., Yang, Y., 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* 14 (1), 1–20.