Computers and Electronics in Agriculture 240 (2026) 111160

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

CRODNet: Cascaded learning network for contactless top-view cattle rotated object detection

Hui Kang [a,1], Yuqi Zhang [a,1], Longxiang Li [a], Chunyang Li [a], Sen Wang [a], Kai Niu [a], Yue Rong [b], Zhiqiang He [a,*]

[a] Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China
[b] School of Electrical Engineering, Computing and Mathematical Sciences Curtin University, Bentley, WA 6102, Australia

ARTICLE INFO

Keywords:
Image alignment
Object detection
Keypoint detection
Precision livestock farming
Cascade learning

ABSTRACT

Cattle recognition and monitoring are critical components of precision livestock farming. However, most existing recognition methods require manual labor and inevitably involve direct contact with cattle, which may cause stress responses in animals. To address these issues, our research focuses on top-view images, with the data acquisition process being contactless. To meet the data requirements for recognition and improve data quality for enhanced recognition accuracy, we partition the overall goal into three sub-tasks, object detection, keypoint detection and image alignment, and designed cattle rotated object detection network (CRODNet) to address them. In the object detection stage, given the inherent characteristics of large body and long rectangular shape, we implemented a pruned multi-head network for rapid cattle pre-selection, ensuring that each image contains a cattle to meet the recognition data requirements. For keypoint detection, we leveraged pre-selection prior and adopted a parallel network architecture with multiple resolutions to extract critical physiological structure features, which are regressed to the cattle's physiological keypoints that include information on body shape and posture. Finally, in the image alignment task, we fully leverage the relationship between the cattle's biological posture and skeletal keypoints from an overhead perspective, modeling a rotation strategy to ensure that cattle with various postures achieve maximum vertical alignment. Experimental results demonstrate that our model achieves a 70% reduction in the number of parameters and 50% decrease in floating point operations (FLOPs), while outperforming other bottom-up approaches by at least 3% in Average Precision (AP). This approach enhances the quality of the pre-selected image and improves the accuracy of the downstream recognition task by at least 2%. These advances are expected to promote the development of precision livestock farming.

1. Introduction

With the rapid advancement of information technology, precision livestock farming has emerged as a transformative approach to revolutionize and enhance the livestock industry (Aquilani et al., 2022). In this context, the ability to identify and track individual animals is crucial for achieving a more refined management system. Continuous monitoring of each animal's health, behavior, and production performance enables the early detection of diseases, the optimization of feeding strategies, and the overall improvement of management practices (Morgan-Davies et al., 2024). However, traditional identification methods (Awad, 2016), such as radio frequency identification (RFID) ear tags, present several challenges. First, since the tags must be applied through direct physical contact, improper application force during this procedure can induce stress responses in animals and pose safety risks to both the animals and handlers. Second, RFID ear tags often necessitate specialized monitoring equipment (Pezzuolo et al., 2020), and the potential for tag loss or failure of associated devices raises both operational costs and data security concerns. Advances in precision livestock farming are now focused on addressing these limitations by offering cost-effective and contactless alternatives for animal monitoring and identification.

Recent advances in artificial intelligence and facial recognition technologies have driven substantial progress in individual animal identification, achieving recognition accuracy rates that exceed 90% in species such as pigs (Wang and Liu, 2022), sheep (Hitelman et al., 2022), and cattle (Ruchay et al., 2024). However, most research has been

* Corresponding author.
E-mail address: hezq@bupt.edu.cn (Z. He).
[1] Hui Kang and Yuqi Zhang contributed equally to this work.

https://doi.org/10.1016/j.compag.2025.111160
Received 3 April 2025; Received in revised form 3 October 2025; Accepted 26 October 2025
Available online 31 October 2025

conducted in controlled environments with pre-selected and aligned images, often requiring manual intervention. Furthermore, some studies have explored the relationship between recognition accuracy and image quality, noting that accuracy can be affected by factors such as posture and facial expressions, which are inherently difficult to control. Although this approach has proven effective in controlled settings, it poses limitations for practical implementation in real-world environments.

To address these challenges, our study focuses on top-view images of cattle. The contactless date collection method requires minimal human intervention: once the equipment is set up, no further interaction with the cattle is necessary, enabling automatic data collection. This approach reduces the potential for inducing stress responses and improves the authenticity and reliability of the data. Although many recent studies (Wang et al., 2024; Ruchay et al., 2020; Lu et al., 2025) have employed 3D data modalities (e.g., LiDAR, depth cameras) for cattle or other livestock modeling and detection, achieving notable progress and potentially providing more accurate representations of body shape and posture, RGB images remain predominant in practical production activities due to their lower cost and greater convenience.

Typically, deep learning-based recognition methods work with images of individual animals. However, in our study, top-view images often contain multiple cattle with varying postures and orientations. Previous research has shown that both image quality (Zhang, 2023) and alignment (Hasan and Pal, 2011) significantly impact recognition accuracy. The primary goal of our study is to extract all the cattle from the image while ensuring high-quality extraction. Traditional object detection methods struggle when dealing with adjacent cattle that have large angular differences, often leading to interference from excessive background or neighboring cattle, which negatively affects recognition accuracy. Therefore, our research focuses on the upstream task of recognition, specifically addressing the challenge of detecting rotated cattle targets under complex conditions, with the aim of aligning the detection results, to ensure that each detected cattle is oriented in a consistent posture, maximizing the proportion of its own information to improve recognition accuracy.

Currently, there are various methods for image alignment, including those based on feature points, regions, and more (Nag, 2017). Considering that pose estimation is also significant for monitoring cattle, by monitoring abnormal behavior or identifying missing keypoints, abnormal and incomplete cow data can be filtered out, in our study, we utilized a keypoint-based approach for image alignment, given that the skeletal keypoints on the cattle's back effectively reflect their posture and body shape. Keypoint detection generally follows two paradigms: top-down and bottom-up approaches. The top-down method involves separate models for object detection and keypoint detection, leading to a more complex and time-consuming workflow. In contrast, the bottom-up method uses a single model to detect keypoints for all targets, first predicting all keypoints and then grouping them by individual targets (Newell et al., 2017). While the bottom-up approach is more efficient, it typically sacrifices some accuracy.

Building upon the methods mentioned above, we designed CRODNet, which cascades object detection and keypoint detection, along with an alignment module. By introducing keypoint detection, we extract physiological structure features from each detected cattle and apply customized adjustments. The alignment module ensures that the detection results for each cattle's posture and orientation are unified, capturing as much of the cattle's intrinsic information as possible, which ultimately enhances recognition accuracy.

In summary, we propose a contactless data collection scheme based on top-view images and construct a corresponding dataset to address the issues inherent in traditional recognition methods. Leveraging deep learning approaches, we decompose the weakly supervised orientation detection task for individuals in varying postures captured in top-view images into three sub-tasks. Specifically designed modules are developed for each sub-task to achieve efficient implementation and integrated into CRODNet, establishing an end-to-end processing flow for top-view images. As our study focuses mainly on image detection and alignment, as well as the impact of alignment on recognition, this approach satisfies the recognition data requirements and improves data quality. The detailed design of each corresponding module is presented as follows.

- Designed object and keypoint detection modules, and cascaded the training of them via a target range pre-aiming (TRPA) module: the object detection module uses a pruned multi-head structure, focusing more on large and long rectangular shape cattle, while the keypoint detection module adopts a multi-resolution parallel structure and uses Squeeze-and-Excitation block, effectively extracting the cattle's physiological structure features. The TRPA module converts global image features into specific target features, enabling the single-target keypoint detection module to effectively handle multi-target image. Through optimization of the structure of these modules, we achieve lightweight designs while improving their performance. Compared to other bottom-up keypoint detection models, CRODNet achieves an improvement of at least 3 AP.
- Designed an alignment module suitable for top-view images: The alignment module leverages the relationship between the skeletal keypoints of the cattle from an overhead perspective, as the stability of keypoints and their connections can indicate the cattle's orientation. Through multi-point joint decision-making to fine-tune alignment parameters such as alignment angle and target range, effectively aligns detected object to improve image quality. Compared to unaligned images, aligned images result in at least a 2% accuracy improvement during the recognition process.

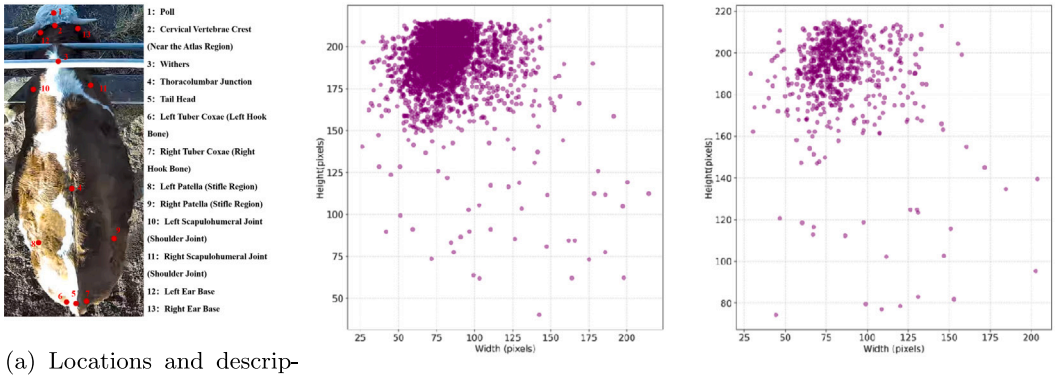
2. Materials and methods

2.1. Data collection

We collected experimental data from a cattle farm in Wuwei, Gansu Province, China, using TPLINK YLIPC-43 A cameras with a 4 mm focal length and a resolution of 2304 × 1296. The dataset consists of beef cattle, including Angus, Simmental, indigenous Yellow cattle, and various crossbreeds derived from these breeds. The cameras were installed 4 m above the cattle feeding channels, with each camera spaced 3 m apart. The overlapping coverage between two adjacent cameras was approximately 20%. Since the left and right areas of cattle within a single shed are not interconnected, we deployed 40 cameras in four areas of cattle within two sheds. This approach, based on the common behavior of cattle during feeding, offers both simplicity and widespread applicability for data acquisition.

A local controller accessed and traversed camera feeds via the Real-Time Streaming Protocol to capture screenshots at three-minute intervals. After daily collection, the local detection module filtered and saved the camera group with the highest number of cattle per frame for upload. This process is contactless and lasts for more than a month, resulting in a total of 874 valid images, nearly 4000 objects, with no cattle were observed in a lateral recumbent position, aside from those that were standing. Each image contains 1 to 6 individuals and was annotated in the COCO dataset format using the COCO-Annotator tool, marking the bounding box and 13 keypoints for each object, as shown in Fig. 1(a). The keypoint annotations followed the guidelines of the Nomina Anatomica Veterinaria (Nomenclature, 2017). These keypoints are crucial for assessing the cattle's posture, body shape, health, and other conditions (Du et al., 2022). Specifically, the keypoints of the head and tail can indicate the length of the cattle, while those of the limbs can provide insight into the width (Yang et al., 2022).

In our study, the images were randomly split into a training set and a test set in a 4:1 ratio and the distribution of cattle sizes is shown in Fig. 1(b). These cattle vary in color, patterns, and body types. Due to



(a) Locations and description of 13 keypoints in cattle.

(b) Distribution of cattle sizes in the training set (left) and test set (right) after resizing the original images to 384×384 pixels

Fig. 1. The distribution of targets' size in the dataset and the annotation descriptions of keypoints.

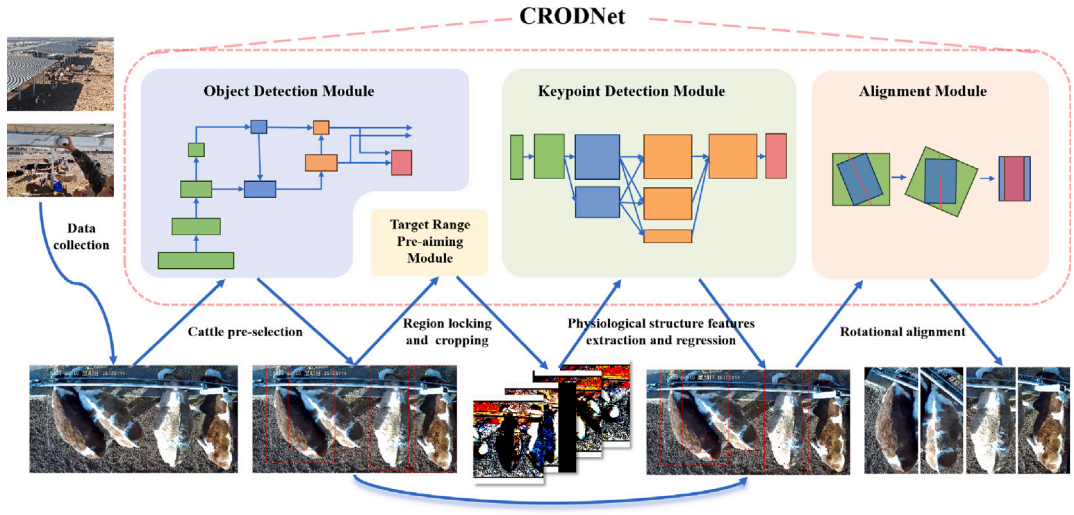


Fig. 2. Processing flow and basic model architecture of the proposed approach.

the contactless method, the acquired images show substantial differences in the number of individuals, their poses, and spatial positions. Furthermore, since the data were collected over a long period, there were variations in lighting conditions and environmental background. These data variations show actual activity patterns of cattle, providing a comprehensive representation of real-world conditions within the target application domain and offering guidance for practical tasks.

To evaluate the model's cross-domain generalization capability, we built an additional dataset at a slaughterhouse using the same acquisition protocol as in the feedlot. Images were captured during the cattle weighing process, where each animal was confined within a pen and photographed overhead. More than 1000 high-resolution images were collected. Compared with the farm, the slaughterhouse differs markedly in background texture, lighting, floor material, and cattle breeds. Nevertheless, the fundamental anatomical structure of the animals remains identical, enabling us to test whether the model truly learns species-level geometric features rather than overfitting to appearance details in the training samples.

2.2. Model structure and workflow

Joint training in a cascade structure, through end-to-end optimization and the use of shared features and additional supervisory signals across stages, may help accelerate convergence and improve performance. This effect has been demonstrated in some studies (Ma et al., 2023; Cai and Vasconcelos, 2018) across diverse domains and datasets, and our task shares certain relevance with these previous works. As shown in Fig. 2, our model employs this approach, where the target range pre-aiming (TRPA) module links object detection and keypoint detection in a cascade. The input images are preprocessed by resizing while maintaining the original aspect ratio. For areas that fall short, pixel padding (114, 114, 114) is applied to ensure uniform image dimensions after processing. During training and inference, the object detection module detects all objects in the image. In the keypoint detection branch, TRPA locks onto target positions and performs region cropping. The keypoint detection module extracts features from the cropped region and predicts the physiological keypoints. The keypoint coordinates are mapped back to the original resolution through inverse operations, by adjusting the coordinates according to the scale factor and the padding applied. Both the keypoint and object information are passed through the alignment module to compute the alignment parameters and perform the final rotational alignment.

2.2.1. Object detection module

The You Only Look Once (YOLO) model (Redmon et al., 2016) uses a single-stage detection method, using multi-scale feature fusion to achieve fast and accurate processing, which has made it widely adopted in the industry. This advantage also aligns well with our task requirements. YOLOv7 features three detection heads designed for large, medium, and small objects, respectively. Identifies 'large' and 'small' objects relative to the input resolution of the network and the

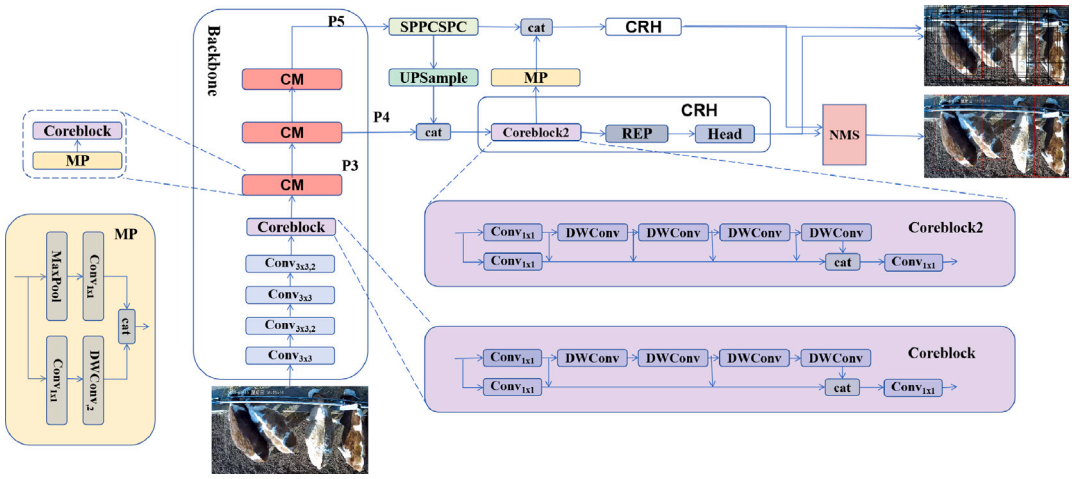


Fig. 3. The structure of dual-head lightweight object detection module. DWConv denotes a 3 × 3 Depthwise Separable convolution, P3, P4, and P5 are the three output branches of the Backbone.

stride of the detection head, based on the relative pixels. In general, cattle on farms are relatively large in size; because we were capturing and processing images of cattle backs from an overhead viewpoint, there is virtually no reduction in target size caused by perspective distortion; the areas of most targets are all on roughly the same scale. The small object detection head is probably not suitable for our task.

As illustrated in Fig. 3, we designed a dual-head lightweight network (DHLN), removing the small detection head by eliminating all components originating from the P3 branch. Since the small detection head focuses on local details using smaller-scale feature maps, it may mistakenly detect areas on the back of cattle as background where feature patterns are similar. This can negatively impact detection performance. While large and medium-sized cattle are easier to detect, and considering the advantages of rapid and cost-effective model training and deployment, we integrated Depthwise Separable Convolutions (DWConv) (Chollet, 2017) to reduce the module's parameters and FLOPs. In the original network, the parameters are primarily influenced by the 3 × 3 convolutions within Efficient Layer Aggregation Networks (ELAN) module (Wang et al., 2023). In our network, the Coreblocks are designed based on the ELAN structure, replacing the standard 3 × 3 convolutions with DWConv, which decomposes standard convolution into depthwise convolution and pointwise convolution. Although pointwise convolution may lose some detailed information, potentially affecting the detection performance, they reduce model parameters and enhance inference speed and computational efficiency, which is beneficial for our task.

We also made necessary modifications to the module's output. This module needs to output all grid-cell predicted bounding box details as well as the final detection results after applying non-maximum suppression (NMS) during training: The former is used for loss calculation and the latter for subsequent keypoint detection.

2.2.2. Target range pre-aiming module

Since our model is jointly trained, after obtaining the bounding box information from the object detection module, we determine the target's position and obtain the target's center coordinates. Then, we set a new fixed-size bounding box (referencing the original HRNet project, with a height of 256 pixels and a width of 192 pixels in our experiments) to ensure consistent input for the keypoint detection module. The target size is smaller than the new bounding box, cattle of all sizes are included and to minimize the impact of potential inaccuracies in early-stage object detection. The x-center of the new bounding box aligns with the target's x-center, and the target's y-center is set to half the height of the original input image. For targets located at the edges of the image, if the new bounding box exceeds the image boundaries, the image will be padded with pixels.

2.2.3. Keypoint detection module

Inspired by the High-Resolution Network (HRNet) (Sun et al., 2019), we designed a compact multi-fusion network (CMFN) tailored for extracting physiological structural features from cattle's back, as shown in Fig. 4. HRNet connects high- and low-resolution convolution streams in parallel, preserving high-resolution representations and effectively fusing features for keypoint detection. However, its multi-branch and multi-stage architecture results in a large number of parameters, which slows down both training and inference. Research (Fan et al., 2023) has shown that excessive stages and branches may not always be beneficial. Therefore, we optimized the structure to better suit the requirements of our task.

We designed the module with three stages, with the third stage as trapezoidal formation using parallel convolutions to reduce redundancy in high-resolution representations. Since inter-stage interactions often involve upsampling and downsampling, which can degrade feature quality, and considering that only high-resolution information is needed for the final layer, we eliminated the final-layer fusion and retained only the first branch, performing only upsampling to preserve high-level feature details.

To enhance the cattle's structural feature extraction ability with fewer convolutions, we integrated Depthwise Over-parameterized Convolution (DOConv) (Cao et al., 2022) into each stage. In the first stage, four bottleneck blocks with residual structure and a 3 × 3 DOConv between 1 × 1 convolutions are employed to increase dimensionality. The second and third stages use basic blocks where traditional 3 × 3 convolutions are replaced with two consecutive 3 × 3 DOConv. DOConv multiplies the parameters of the depthwise and regular convolution kernels, boosting feature extraction with a minimal parameter increase. Moreover, to address the potential introduction of redundant information from multi-branch fusion, we added a lightweight Squeeze-and-Excitation (SE) module (Hu et al., 2018) after the DOConv operations in basic block. This module adaptively re-weights channel features, emphasizing the most informative cues related to the cattle's back physiological structure while suppressing redundant ones, making the detected keypoints more precise.

2.2.4. Alignment module

Considering that our task involves aligning all targets into a uniform posture and orientation and that the livestock are typically in a feeding state during photo capture, the general posture variations manifest mainly in the head direction and slight head turns, while the back remains relatively uniform. Even with some twisting or bending, the range of motion is limited due to physiological constraint, especially in larger individuals, resulting in minimal impact. Therefore, we use

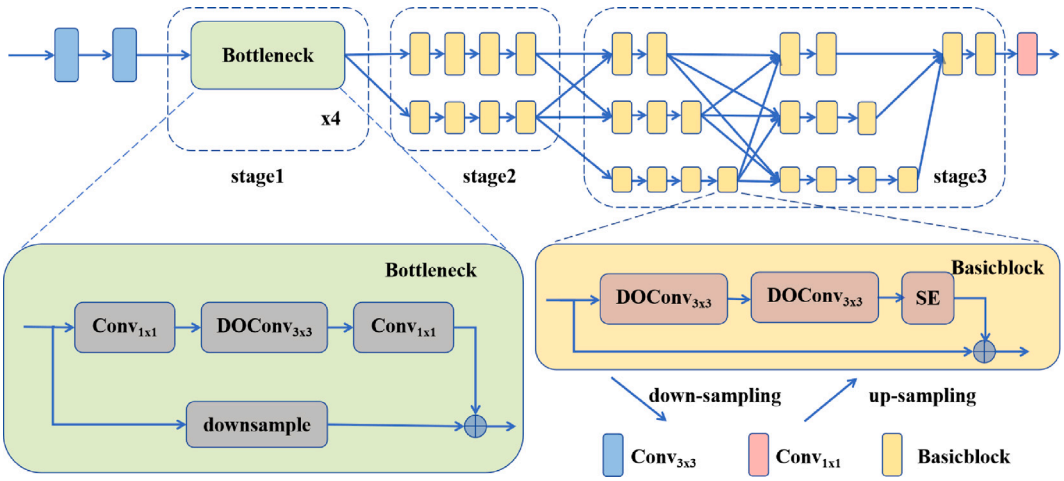


Fig. 4. The structure diagram of compact multi-fusion keypoint detection module.

the back as a reference to minimize the influence of head posture information as much as possible.

For these relatively fixed dorsal keypoints, we calculate the corresponding sequential offset angles and assign higher weights based on physiological and positional constraints, such as the relatively stable positions of the upper and lower limbs. While head posture varies more, its influence is relatively limited due to the smaller size compared to the body. Therefore, we introduce neck keypoints with lower weights to adjust the offset angle computations, which is especially beneficial for cattle in extreme postures. In cases where some keypoints are missing due to image capture issues, the weights for the corresponding angles are set to zero. Ultimately, these angles are aggregated through weighted summation to determine the optimal rotation angle for alignment. The formula for the rotation angle Θ is as follows:

$$\Theta = \frac{\sum_{i=1}^{4} W_i \theta_i}{\sum_{i=1}^{4} W_i} \tag{1}$$

where: θ_i represents the angle formed between the specified keypoint line segments and the y-axis. W_i is the weight assigned to each angle. Considering that the limbs are relatively more stable than the head, we assign higher weights to the limbs. In our experiment, the weights are set as 4, 4, 1, 1 for the limbs and the head, respectively.

Because alignment reduces redundant background pixels and increases the proportion of bovine anatomical information, the aligned box no longer overlaps with the original detection box. Considering factors such as cattle posture, intra-species structural uniformity, and interference from adjacent cattle, we developed an algorithm to compute the aligned box. The height h is obtained as:

$$h = \max_{y1 \in Y_{\text{tail}}} y1 - \min_{y2 \in Y_{\text{head}}} y2 + offset \tag{2}$$

Here, Y_{tail} and Y_{head} represent the set of y coordinates of the keypoints related to the tail and head, while *offset* compensates for the residual distance between the head/tail of cattle and the physical ends of the torso, and the value of *offset* is calculated by determining the mapped distance along the vertical axis for the corresponding angle, based on the distances from the head and tail keypoints to the top and bottom edges of the detection box.

The width w jointly considers body uniformity, posture variations and adjacent cattle overlap:

$$w = g(\mathbf{k})\, f(h) - \frac{m}{2} \tag{3}$$

where $f(h)$ converts the estimated height into an initial width. Under normal circumstances, cattle have relatively uniform body shapes. We analyzed the distribution of the width–height ratio for cattle whose rotation angles deviate by less than 3°; the center of this distribution is treated as the initial ratio and multiplied by h to obtain a preliminary width. Because highly curved postures lead to larger width–height ratios, we introduce a correction factor $g(\mathbf{k})$ (k denotes all keypoints), estimated from the relative locations of keypoints to penalize severe bending, the more the posture is bent, the larger the value. Finally, if the aligned boxes of two adjacent cattle overlap, each box is trimmed inwards by half of the overlap (m) to avoid mutual interference.

Through the above rotation-alignment procedure, we obtain single-cattle images that are orientation consistent and contain markedly less background clutter (Fig. 5 for a visual comparison before and after alignment).

3. Experiment setup

3.1. Experimental environment

This project was developed on a Linux operating system using the PyTorch deep learning framework, implemented in Python. The PyTorch version used is 1.10.0, and the Python version is 3.8. All experiments were conducted on a GPU server to accelerate the training process.

3.2. Training configuration

During the experiments, the random seed was 2, the initial learning rate was 0.001, the batchsize was 4 and the optimizer used was AdamW (Loshchilov and Hutter, 2019) with weight-decay of 0.0001. Training was conducted over 130 epochs, with the learning rate multiplied by 0.05 at the 60th and 90th epochs. Prior to training, data augmentation was employed to expand the dataset and enhance the model's generalization capabilities. Data augmentation techniques included image scaling, adjustment of lighting intensity and contrast, image translation, and horizontal flipping, with augmentation intensities maintained within reasonable ranges (Shorten and Khoshgoftaar, 2019). Each other model was trained according to the original official configuration files and loaded with pre-trained models to accelerate training. The pre-trained keypoint detection model was trained on the COCO 2017 dataset. YOLOv7 and HRNet-W48 were used as the experimental and comparative models for detection.

Since this study involves joint training, the loss function employs a weighted summation approach that combines the object detection and keypoint detection loss. Specifically, object detection utilizes Complete Intersection over Union (CIoU) loss (Zheng et al., 2021), while keypoint detection uses the Mean Squared Error (MSE). As both values stabilize at the same order of magnitude, the weights were set in a 1:1 ratio. However, the keypoint detection loss was relatively large in the first

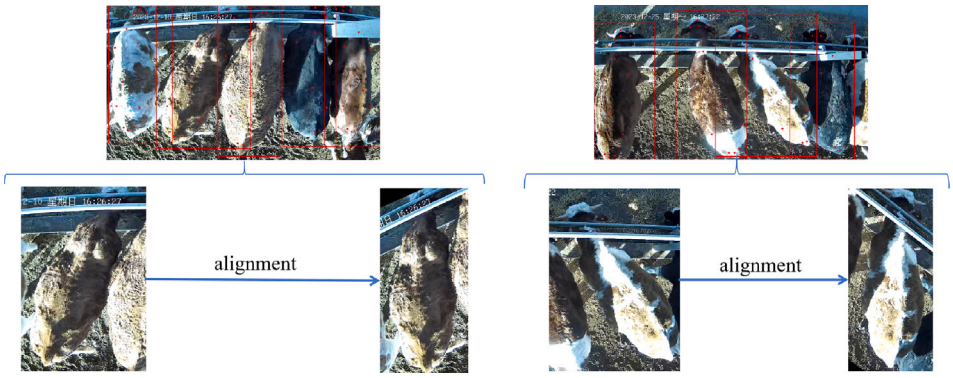


Fig. 5. The results before and after our alignment algorithm.

few epochs. To ensure proper parameter updates, if the keypoint detection loss exceeded 20 in the first few epochs, it was reduced by a factor of 10.

3.3. Evaluation metrics

Considering that the ground-truth keypoints in data annotations may contain certain errors, we adopted the Object Keypoint Similarity (OKS) method to calculate the prediction accuracy of keypoint detection.

$$\text{OKS} = \frac{\sum_i \exp\left(-\frac{d_i^2}{2s^2\sigma_i^2}\right) \cdot \delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{4}$$

where:

- i represents the index of the keypoint, $i = 1, 2, \ldots, 13$.
- d_i is the Euclidean distance between the predicted position and the ground truth position of the ith keypoint.
- s is the scale of the object, typically taken as the square root of the area of the target's bounding box.
- σ_i is the normalization factor for the ith keypoint, used to adjust the tolerance for different keypoints. During the evaluation of the metrics, all the sigma values were set to 0.04.
- $\delta(v_i > 0)$ is the indicator function, which takes the value 1 when the ith keypoint is visible in the ground truth annotation ($v_i > 0$), and 0 otherwise.

We utilize Average Precision (AP) as the primary performance metrics for evaluating models in detection task. Additionally, we consider AP at specific OKS or Intersection over Union (IoU) thresholds, AP@0.5 (OKS) for keypoint detection and AP@0.5 (IoU) for object detection, to provide more granular insights into model performance. In our experiments, AP refers to AP@0.5–0.95, averaged over thresholds from 0.5 to 0.95 in steps of 0.05. The recall score (AR) is included as an additional performance metric. These metrics were calculated using the official COCO evaluation API. The recognition task uses Accuracy as the evaluation metric. Accuracy is defined as the proportion of correctly predicted positive and negative samples to the total number of samples. The results report the 95% confidence intervals (CI) after multiple runs, represented in tables as (±).

4. Experiment results

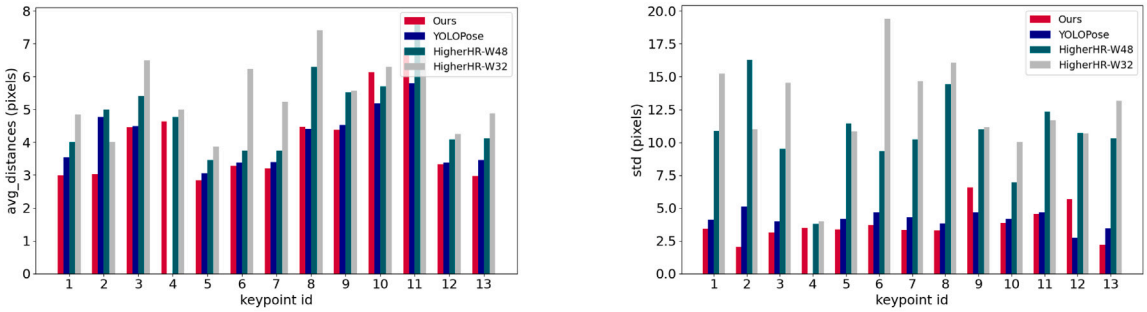
4.1. Comparison experiments

To comprehensively evaluate the performance of our model, we compared it with models referenced in the literature. Since our model primarily addresses the accuracy shortcomings of bottom-up models and the slower running speed of top-down models, the comparison was divided into two groups:

Bottom-up Models: Including YOLOPose (Maji et al., 2022) and HigherHRNet (Cheng et al., 2020), mainly compared the keypoint detection accuracy, as shown in Table 1. Compared to classical bottom-up models, our model achieved an improvement of at least 3% in AP. Although HigherHRNet-W32 has the fewest parameters, its accuracy is relatively lower. HigherHRNet-W48 achieves higher accuracy by utilizing more channels, but this comes with a substantial increase in parameters, and its AP still lower than our model by 4.0%. In the case of the YOLOPose model, an accuracy improvement of 9.7% was achieved by excluding the fourth keypoint, as the model consistently predicted the y-coordinate of this keypoint as 0, leading to a significant deviation from the actual coordinates. If we relaxed the limitation for this keypoint and set its sigma to 0.5, the overall AP would increase by 9.7%. Nevertheless, our model still achieves competitive accuracy with relatively few parameters.

Top-down Models: Such as YOLO combined with HRNet (Nguyen et al., 2022), mainly compared the inference time, parameters, FPS (Frames Per Second), FLOPs and memory usage, as shown in Table 2. For inference time, we evaluated the time required by our proposed alignment algorithm to process the entire test set, measured in seconds. The testing involved object detection, keypoint detection, and all post-processing steps, including alignment, with the corresponding time for each step. As demonstrated by the results, our model achieves a more than 70% reduction in parameters, over a 50% decrease in FLOPs, and nearly a 20% reduction in inference time compared to the top-down approach. Since keypoint detection is performed on each detected cattle in object detection, the processed images are several times greater than that of object detection, and as object detection accounts for the model inference and NMS time, keypoint detection and post-processing consume more time. The post-processing itself takes a considerable amount of time, which makes the overall process have a relatively low FPS. However, this is sufficient for our task, and the results show that our improvements enhance operational efficiency. Memory usage reflects the maximum GPU memory allocation during model inference, calculated using PyTorch's built-in library functions. Despite a reduction in the number of parameters, memory usage has increased due to our model's parallel processing approach, where the original data is retained and intermediate variables are processed on the GPU. However, when the modules are run in the same top-down method, memory usage is relatively lower.

As shown in Fig. 6, we calculated and plotted the average distance error and the standard deviation of the prediction distance for 13 keypoints for each model to provide a more intuitive comparison of keypoint detection performance between different models. The distance error is obtained by calculating the pixel distance between the ground truth and the predicted keypoints. As shown, our model exhibits smaller deviations in predictions for most keypoints compared to other models



(a) Distribution of average distance error (b) Distribution of standard deviation

Fig. 6. Distribution of both average and standard distance errors of 13 keypoints predicted by different models.

Table 1
Comparison of the proposed model's parameters and AP with other models.

Models	AP	AP@0.5	Parameter (M)
HigherHRNet-W32	0.713	0.943	28.64
HigherHRNet-W48	0.731	0.951	63.82
YOLOPose	0.644(+0.097)	0.974	80.15
Ours	0.771	0.988	30.46

and demonstrates more stable predictions, with both the average distance error and standard deviation generally lower than those of other models, especially compared to HigherHRNet.

To further quantify how prediction errors propagate to the rotation angle Θ, we performed a Monte-Carlo sensitivity analysis based on the error statistics. Starting from the ground-truth annotations, we injected Gaussian noise whose μ, σ matched the per-keypoint prediction error in Fig. 6, and drew 1 000 000 samples for each configuration. For each sample, we recomputed Θ and recorded (ΔMean (°)) the absolute deviation from the ground-truth, (Std. Dev. (°)) the sample-to-sample standard deviation of Θ, and the offsets of the quartiles and CI relative to the median (ΔQuartiles (°) and Δ95% CI (°)).

We examined three perturbation settings: (i) noise on keypoint 4 only, (ii) simultaneous noise on keypoints 3–5, and (iii) noise on all keypoints. For each setting, we evaluated two noise levels, $1 \times (\mu, \sigma)$ and $2 \times (\mu, \sigma)$, and repeated the full simulation ten times to obtain stable estimates. The results are reported in Table 3. Adding noise to a single keypoint or a small subset of keypoints produces only minute deviations in the estimated rotation angle, with the error confined to a narrow band that remains limited even when the noise level is doubled. Injecting noise into all keypoints naturally increases the variation, yet the standard deviation of Θ never exceeds three degrees and stays below six degrees even under the doubled-noise regime. And from the quartiles and 95% CI, it can be observed that the angle distribution is relatively concentrated. A three-degrees rotation is practically imperceptible in the final image, the impact on alignment is negligible. It to some extent proves weakly sensitive to keypoints prediction errors and shows the stability of our method.

4.2. Ablation experiments

To evaluate the individual effects of improvements in each module, we conducted a series of experiments on our dataset. Table 4 shows the effects of various modifications in the object detection module, Baseline stand for YOLOv7, Two Head stand for reducing the number of detection heads to two, and DWConv stand for introducing DWConv in the CoreBlock and MP module. Table 5 shows the results of keypoint detection module, Baseline stand for HRNet, No mix stand for abandoning feature fusion at the last input layer, and SE stand for SE module at the end of basic block. Each experiment builds on the previous one to evaluate the impact of each improvement. In the object detection module, removing the small object detection head enhances model suitability for our task, resulting in improved AP. Introducing DWConv improves performance over the baseline while reducing parameters and FLOPs. In the keypoint detection module, the modified three-stage structure reduces parameters and FLOPs. Eliminating feature fusion in the final layer mitigates information loss from feature upsampling, improving AP. The introduction of DOConv enhances feature extraction capabilities by slightly increasing the number of learnable parameters, further enhancing AP. Lastly, incorporating SE modules at the end of basic blocks improves adaptability, leading to additional gains in AP.

We also conducted experiments with images of different resolutions as input to test the robustness of removing the P3 detection head in the object detection module. Table 6 shows the impact of retaining versus removing the P3 detection head on object detection performance. In all cases with different resolutions, removing the P3 detection head led to a general improvement in AP. The results indicate that the distinction of small objects is more dependent on their relative size than on absolute pixels and image resolutions, removing the P3 detection head may reduce the interference of redundant background information, which is beneficial for our task.

As shown in Table 7, we conducted a series of experiments to assess the impact of each module within our model. Each set of experiments involved replacing specific modules within the model. Since the AP of the two tasks is on the same order of magnitude, we calculate the average AP to provide a more intuitive comparison of the model's overall performance across these two tasks.

The experimental results demonstrate that individual replacement of the object detection module with DHLN could improve object detection AP. Although the keypoint detection AP slightly decreases under joint training, the model's parameters and FLOPs are reduced, and the average AP experiences a substantial increase. Similarly, individually replacing the keypoint detection module with CMFN, its AP remains stable without decline, the object detection AP achieves a certain improvement, and the model's parameters and FLOPs are greatly reduced. When both modules are replaced with the optimized network, both object detection and keypoint detection AP are enhanced under joint training, along with reductions in parameters and FLOPs. These findings indicate that our improvements are advantageous for our task and beneficial for joint training.

To further investigate the generalization ability of the model and compare it with the baseline, we incorporated the slaughterhouse dataset and designed two experiments: (1) training on the farm dataset while testing on slaughterhouse dataset; (2) mixing the training portions of the farm and slaughterhouse datasets at a 1:1 ratio (resulting in a combined training set of 2400 cattle that is smaller than the original farm training set) and validating on the farm test set with a 4:1 split between training and testing cattle. Although the two domains differ significantly in surroundings, cattle categories and surface appearance, they share the same underlying bovine anatomy and physiological structure. As shown in Table 8, despite cross-dataset

Table 2
Comparison results of the proposed model with the top-down method.

Methods	Parameter (M)	FLOPs (G)	Time (s)	FPS	Memory (MB)
YOLO+HRNet	100.79	66.41	89.35 (3.14 + 56.13 + 30.08)	1.98	933.15
Ours	30.46	29.49	71.40 (2.76 + 38.34 + 30.30)	2.48	1221.93 (554.07)

Table 3
Monte-Carlo sensitivity analysis of the rotation Θ with respect to keypoints prediction errors.

Injected noise		ΔMean (°)	Std. Dev. (°)	ΔQuartiles (°)	Δ95% CI (°)
Keypoint 4	$1\times(\mu,\sigma)$	0 (±0.001)	0.298 (±0.001)	0.16 (±0.001)	0.633 (±0.001)
	$2\times(\mu,\sigma)$	0.001 (±0.002)	0.598 (±0.001)	0.33 (±0.001)	1.257 (±0.002)
Keypoints 3–5	$1\times(\mu,\sigma)$	0 (±0.002)	0.633 (±0.002)	0.42 (±0.001)	1.254 (±0.002)
	$2\times(\mu,\sigma)$	0.002 (±0.001)	1.277 (±0.002)	0.84 (±0.002)	2.519 (±0.002)
All keypoints	$1\times(\mu,\sigma)$	0 (±0.005)	2.512 (±0.006)	1.67 (±0.005)	4.946 (±0.006)
	$2\times(\mu,\sigma)$	0.001 (±0.011)	5.081 (±0.011)	3.38 (±0.0012)	10.058 (±0.011)

Table 4
Ablation experiments of the object detection module, with the table demonstrating the impact of different improvements on the module's performance.

No.	Baseline	Two head	DWConv	AP	Recall	Parameter (M)	FLOPs (G)
1	✓			0.773	0.918	37.20	18.95
2	✓	✓		0.788	0.922	26.87	14.95
3	✓	✓	✓	0.782	0.921	17.49	9.65

Table 5
Ablation experiments of the keypoint detection module, with the table demonstrating the impact of different improvements on the module's performance.

No.	Baseline	Three stage	No mix	DOConv	SE	AP	AR	Parameter (M)	FLOPs (G)
1	✓					0.756	0.798	63.60	47.46
2	✓	✓				0.773	0.811	15.71	22.74
3	✓	✓	✓			0.777	0.816	12.54	19.74
4	✓	✓	✓	✓		0.781	0.820	13.23	19.83
5	✓	✓	✓	✓	✓	0.787	0.825	13.30	19.84

Table 6
Object detection AP of retaining versus removing the P3 head with different resolutions.

Resolution	With P3 head	Without P3 head
640 × 640	0.798 (±0.001)	0.805 (±0.002)
480 × 480	0.786 (±0.001)	0.800 (±0.001)
320 × 320	0.767 (±0.001)	0.784 (±0.002)

training and testing, which reduced keypoint detection AP by 0.1, our method still achieves an AP near 0.70, with AP@0.5 remaining virtually unchanged. Compared to HRNet, our model performs better, demonstrating a stronger generalization capability. In the other case, diluting the farm data during training by mixing an equal proportion of entirely independent slaughterhouse samples had little impact on the model's test performance. The results indicate that our model captures intrinsic inter-sample regularities and the spatial distribution of physiological structures, rather than simply memorizing the data distribution.

4.3. Alignment algorithm experiments

To evaluate the impact of alignment operations on recognition accuracy, we restructured the dataset assigning different individuals to different categories. To reduce the impact of the number of samples on recognition accuracy, we randomly selected five images per class from the training set as training samples. We then conducted experiments with and without the alignment operation. To further validate the generalizability of the algorithm's effect on recognition accuracy, we also conducted corresponding experiments on slaughterhouse images and the publicly available OpenCows2020 dataset (Dataset Ninja, 2025). The OpenCows dataset contains numerous images per class for recognition task, if all samples from the training set were used for training, the recognition accuracy will stably exceed 99%. Therefore, we randomly selected five images from each class as training samples and repeated this process five times. Under identical conditions, we trained and tested the Omni-Scale Network (OSNet) (Zhou et al., 2019) with pre-trained parameters and computed the recognition accuracy. The results are presented in Table 9. Since the OpenCows dataset provides a specific split for unknown class testing to evaluate the model's ability to recognize new classes, we conducted comparative experiments according to this division (Table 10). Columns 1–5 represent the results of five random splits of the training data and columns labeled 10:90, 20:80, etc. indicate the ratio of known to unknown classes. Similarly, as shown in Table 9, in our dataset we took the slaughterhouse data as the training set and the farm data as the test set, and compared aligned and non-aligned to evaluate their impact on recognition of unknown classes.

The experimental results indicate that alignment can generally improve recognition accuracy as the alignment process effectively reduces redundant information in the images and increases the proportion of biological information relevant to the cattle. Although the improvement on the OpenCows dataset is not significant, the method remains effective in few-shot scenarios. Moreover, for unknown class recognition, alignment achieves notably better results in certain cases, such as with 20:80 and 30:70 known-to-unknown class splits. Given the constraints of the slaughterhouse environment, the slaughterhouse dataset has only three interval records per cattle, and some only have two. Due to the

Table 7
The table presents the ablation experimental results of the object detection module and the keypoint detection module on the overall model's performance.

Object & Keypoint detection modules	Object detection AP	Keypoint detection AP	Average AP	Parameter (M)	FLOPs (G)
YOLO+HRNet	0.768 (±0.008)	0.767 (±0.001)	0.768 (±0.004)	100.79	66.41
DHLN+HRNet	0.816 (±0.004)	0.764 (±0.004)	0.791 (±0.002)	81.09	55.21
YOLO+CMFN	0.785 (±0.008)	0.767 (±0.001)	0.775(±0.004)	50.17	38.79
Ours	0.817 (±0.007)	0.771 (±0.002)	0.794 (±0.003)	30.46	29.49

Table 8
Cross-domain keypoint detection results (new dataset training and mixed dataset training).

Methods	AP	AP@0.5	AR
HRNet (new)	0.660 (±0.011)	0.939 (±0.006)	0.720 (±0.013)
Ours (new)	0.674 (±0.009)	0.938 (±0.007)	0.722 (±0.010)
HRNet (mixed)	0.737 (±0.012)	0.980 (±0.005)	0.782 (±0.009)
Ours (mixed)	0.763 (±0.003)	0.982 (±0.007)	0.806 (±0.003)

Table 9
Cattle recognition accuracy in our dataset and OpenCows with and without alignment, and unknown-class testing (UKC-Testing) in our dataset.

Method	Accuracy
Ours with alignment	95.92% (±0.85)
Ours without alignment	92.61% (±0.99)
OpenCows with alignment	91.37% (±0.45)
OpenCows without alignment	90.55% (±0.56)
UKC-Testing with alignment	79.90% (±0.96)
UKC-Testing without alignment	77.90% (±1.13)

setting and evaluation of few shots in unseen samples, the recognition accuracy is relatively low; nevertheless, the model still achieves 80% accuracy, and through alignment, the accuracy can be improved by 2%, demonstrating good generalization performance.

4.4. Result visualization and discussion

We visualized the output of the model using a heatmap to analyze the model's prediction ability of keypoints, as shown in Fig. 7. The intensity of the heatmap colors represents the confidence of the model in detecting a keypoint at a specific location. The brighter the color (such as red or yellow), the higher the probability that the model predicts in that region. It can be seen that the focus of the model aligns closely with the true distribution of keypoints on cattle. However, due to the larger area corresponding to the limbs, predicting in these regions is more challenging, resulting in lower confidence in these areas on the heatmap. This observation is consistent with the previous results of the keypoints' prediction error analysis.

To demonstrate the actual performance of the model, we tested our model on the test set and visualized the results. The experiment considered both objective factors, such as lighting and weather conditions like snow, which interfere with the original images, and subjective factors, including shadows and varying postures caused by the dense arrangement of cattle during feeding. In Fig. 8, (a) and (d) demonstrate that the model achieves good detection results under varying lighting conditions; (c) shows that the model's performance is slightly impacted, even when there are significant color differences among individuals in the image; The comparison between (a) and (b) illustrates that the level of interference between individuals, whether sparse or crowded, has little impact on the model's predictions; Additionally, (c) and (e) indicate that even when the back is covered with a thin layer of snow, blending with the background and causing interference, the predictions remain relatively accurate. We also visualized the predictions of the comparative bottom-up models, as shown in Fig. 9. It is evident that the HR-model exhibits large errors in certain keypoints, as shown in the red box at the bottom of (a), there were originally supposed to be only three

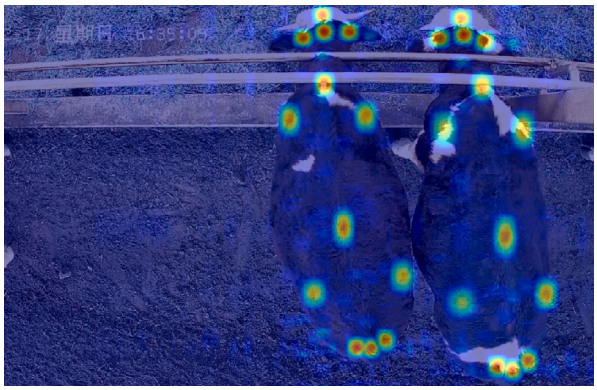


Fig. 7. Heatmap of the model's keypoint detection output.

keypoints, but additional keypoints were also predicted in this area. The YOLOPose model demonstrates almost no predictive capability for keypoint 4, as shown in the red box in the middle of (b), keypoint 4 was supposed to be present but was not predicted at all. However, these less accurate predictions commonly occur when the head region is heavily occluded, when individuals are positioned too closely together, or when the posture is severely distorted, as shown in the top box of each image. We attribute these inaccuracies to limitations in the dataset, as such extreme cases are underrepresented during training.

Many existing studies on cattle re-identification have validated the effectiveness of our cattle back image data processing approach, and several datasets (Andrew et al., 2021; Sharma et al., 2025) have been developed focusing on cattle back images for re-identification tasks, aiming to enable hands-off cattle tracking and monitoring applications in precision farming. Although factors such as the pattern on the cattle's back and their different growth stages can affect recognition, and pure-colored cattle might pose more identification challenges, this research primarily focuses on image detection and alignment, as well as the impact of alignment on recognition. The influence of cattle's biometric features on recognition is not the main focus of our study. Since the physiological structure of cattle is similar across different breeds, their growth stages and varying back patterns do not substantially affect the distribution of keypoints, which is unlikely to significantly impact the performance of our detection and alignment model. And we will investigate their impact more thoroughly in future work.

Although our proposed approach and model demonstrate good performance in this type of data, we have noticed some potential limitations. Specifically, the model exhibits reduced prediction accuracy in scenarios where the cattle were severely obscured, which may be primarily due to the limited training samples in such conditions. In addition, field conditions like adverse weather and poor visibility, as well as practical constraints from equipment and environment on cattle farms can influence the performance of data collection and equipment. More extreme conditions, such as heavy rain, dense fog, low illumination at night (without auxiliary lighting), or thick mud covering the back, together with long-term degradation factors like slight camera displacement or vibration, may introduce noise and reduce image quality. Moreover, low-cost cameras and related devices with limited

Table 10
Known-class training and unknown-class testing accuracy.

	10:90	20:80	30:70	40:60	50:50
Without align	70.59 (±1.44)	73.69 (±1.98)	81.95 (±0.30)	87.98 (±0.46)	90.74 (±0.84)
With align	70.62 (±1.11)	75.44 (±1.23)	84.66 (±0.27)	89.11 (±0.58)	90.90 (±1.01)

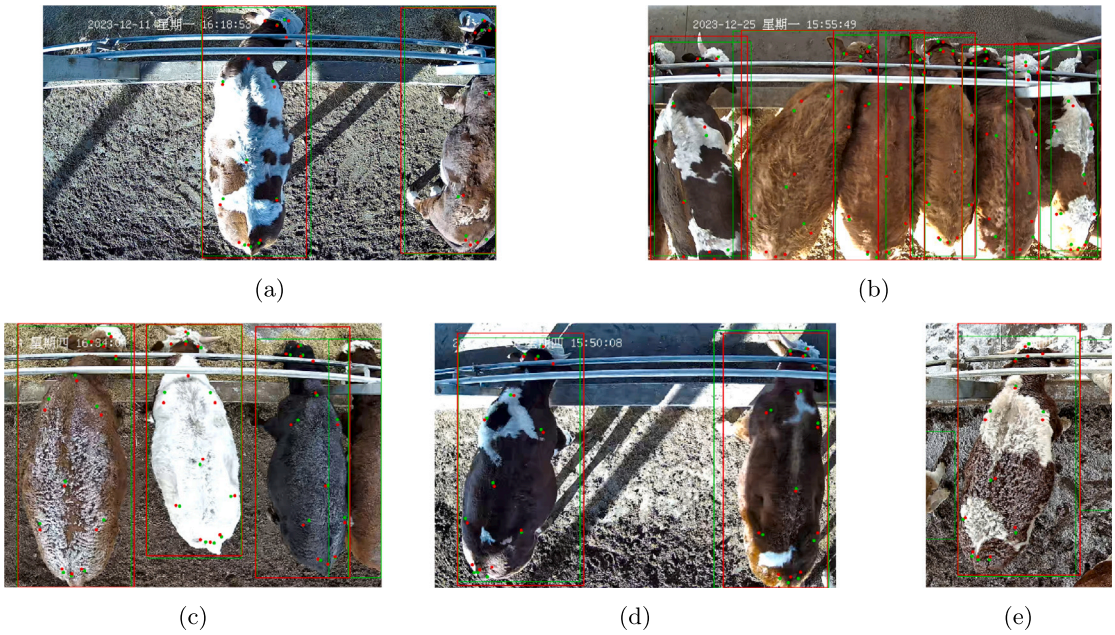


Fig. 8. The group of figures illustrates the result of our model in different situations. The red annotations represent the predicted results, while the green annotations represent the ground truth.

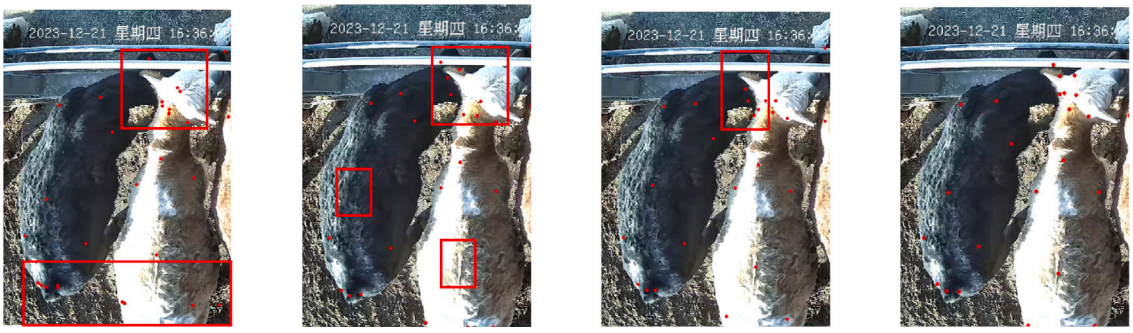


Fig. 9. Comparison of results across different models. (d) represents the ground truth annotations. Red bounding boxes delineate regions with significant deviations from ground truth annotations.

performance may be more susceptible to interference or damage, requiring more robust, higher-performance hardware. However, since our data undergo automatic filtering through a detection model after collection to discard unusable or heavily noisy images, the impact of these extreme factors will be reduced.

5. Conclusion

This paper introduces a data processing solution designed for contactless top-view images of cattle. The newly designed cascade model, along with the multi-point decision-based alignment algorithm, enhances detection performance while maintaining a lightweight design, which is also beneficial for downstream recognition tasks. Additionally, due to the relatively similar physiological structures and keypoint distribution patterns of the backs of large livestock, our top-view image acquisition and alignment scheme may potentially be adapted for other livestock such as sheep and pigs. This low-cost and convenient data collection solution, combined with a lightweight and efficient model, may provide new ideas and solutions for recognition, data acquisition and processing, and may help advance the development of precision livestock farming.

In future work, we will conduct further experiments on other livestock based on physiological structural similarities to validate the effectiveness of our method and explore the inherent relationships between these animals. Additionally, we will extend the current research on recognition, investigating more factors that affect recognition in

extreme or open environments. Ultimately, if conditions permit, we will consider incorporating updated data types (3D) and alignment methods through point cloud in future research, and aim to integrate detection and recognition, applying it to practical production activities to promote the development of precision livestock farming.

CRediT authorship contribution statement

Hui Kang: Writing – original draft, Validation, Project administration, Methodology, Investigation, Formal analysis. **Yuqi Zhang:** Writing – review & editing, Resources, Project administration, Methodology, Investigation. **Longxiang Li:** Writing – review & editing, Methodology, Investigation, Data curation. **Chunyang Li:** Resources, Data curation. **Sen Wang:** Resources, Data curation. **Kai Niu:** Writing – review & editing, Methodology, Formal analysis. **Yue Rong:** Resources. **Zhiqiang He:** Writing – review & editing, Supervision, Methodology, Formal analysis.

Declaration of competing interest

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

Acknowledgment

This paper was supported by the BUPT innovation and entrepreneurship support program (project funding number: 2025-YC-S004).

Data availability

Data will be made available on request.

References

Andrew, William, Gao, Jing, Mullan, Siobhan, Campbell, Neill, Dowsey, Andrew W., Burghardt, Tilo, 2021. Visual identification of individual Holstein-Friesian cattle via deep metric learning. Comput. Electron. Agric. 185, 106133.

Aquilani, C., Confessore, A., Bozzi, R., Sirtori, F., Pugliese, C., 2022. Review: Precision livestock farming technologies in pasture-based livestock systems. Animal 16 (1), 100429.

Awad, Ali Ismail, 2016. From classical methods to animal biometrics: A review on cattle identification and tracking. Comput. Electron. Agric. 123, 423–435.

Cai, Zhaowei, Vasconcelos, Nuno, 2018. Cascade R-CNN: Delving into high quality object detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6154–6162.

Cao, Jinming, Li, Yangyan, Sun, Mingchao, Chen, Ying, Lischinski, Dani, Cohen-Or, Daniel, Chen, Baoquan, Tu, Changhe, 2022. DO-conv: Depthwise over-parameterized convolutional layer. IEEE Trans. Image Process. 31, 3726–3736.

Cheng, Bowen, Xiao, Bin, Wang, Jingdong, Shi, Honghui, Huang, Thomas S., Zhang, Lei, 2020. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5385–5394.

Chollet, François, 2017. Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1800–1807.

Dataset Ninja, 2025. Visualization tools for OpenCow2020 dataset. visited on 2025-06-10. https://datasetninja.com/opencows2020.

Du, Ao, Guo, Hao, Lu, Jie, Su, Yang, Ma, Qin, Ruchay, Alexey, Marinello, Francesco, Pezzuolo, Andrea, 2022. Automatic livestock body measurement based on keypoint detection with multiple depth cameras. Comput. Electron. Agric. 198, 107059.

Fan, Qingcheng, Liu, Sicong, Li, Shuqin, Zhao, Chunjiang, 2023. Bottom-up cattle pose estimation via concise multi-branch network. Comput. Electron. Agric. 211, 107945.

Hasan, Md. Kamrul, Pal, Christopher J., 2011. Improving alignment of faces for recognition. In: 2011 IEEE International Symposium on Robotic and Sensors Environments. ROSE, pp. 249–254.

Hitelman, Almog, Edan, Yael, Godo, Assaf, Berenstein, Ron, Lepar, Joseph, Halachmi, Ilan, 2022. Biometric identification of sheep via a machine-vision system. Comput. Electron. Agric. 194, 106713.

Hu, Jie, Shen, Li, Sun, Gang, 2018. Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Loshchilov, Ilya, Hutter, Frank, 2019. Decoupled weight decay regularization. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, la, USA, May 6-9, 2019.

Lu, Hexiao, Zhang, Jialong, Yuan, Xufeng, Lv, Jihong, Zeng, Zhiwei, Guo, Hao, Ruchay, Alexey, 2025. Automatic coarse-to-fine method for cattle body measurement based on improved GCN and 3D parametric model. Comput. Electron. Agric. 231, 110017.

Ma, Shuailei, Wang, Yuefeng, Wei, Ying, Fan, Jiaqi, Li, Thomas H., Liu, Hongli, Lv, Fanbing, 2023. CAT: LoCalization and IdentificAtion cascade detection transformer for open-world object detection. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 19681–19690.

Maji, Debapriya, Nagori, Soyeb, Mathew, Manu, Poddar, Deepak, 2022. YOLO-pose: Enhancing YOLO for multi person pose estimation using object keypoint similarity loss. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 2636–2645.

Morgan-Davies, C., Tesnière, G., Gautier, J.M., Jørgensen, G.H.M., González-García, E., Patsios, S.I., Sossidou, E.N., Keady, T.W.J., McClearn, B., Kenyon, F., Caja, G., Grøva, L., Decandia, M., Cziszter, L., Halachmi, I., Dwyer, C.M., 2024. Review: Exploring the use of precision livestock farming for small ruminant welfare management. Animal 18, 101233, Selected keynote lectures of the 74th Annual Meeting of the European Federation of Animal Science (Lyon, France).

Nag, Sayan, 2017. Image registration techniques: A survey. ArXiv, abs/1712.07540.

Newell, Alejandro, Huang, Zhiao, Deng, Jia, 2017. Associative embedding: End-to-end learning for joint detection and grouping. In: Guyon, I., Luxburg, U. Von, Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc.

Nguyen, Hung-Cuong, Nguyen, Thi-Hao, Nowak, Jakub, Byrski, Aleksander, Siwocha, Agnieszka, Le, Van-Hung, 2022. Combined YOLOv5 and HRNet for high accuracy 2D keypoint and human pose estimation. J. Artif. Intell. Soft Comput. Res. 12, 281–298.

Nomenclature, International Committee On Veterinary Gross Anatomical, 2017. Nomina Anatomica Veterinaria, sixth ed. World Association of Veterinary Anatomists.

Pezzuolo, Andrea, Guo, Hao, Guercini, Stefano, Marinello, Francesco, 2020. Non-contact feed weight estimation by RFID technology in cow-feed alley. In: 2020 IEEE International Workshop on Metrology for Agriculture and Forestry. MetroAgriFor, pp. 170–174.

Redmon, Joseph, Divvala, Santosh, Girshick, Ross, Farhadi, Ali, 2016. You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, Los Alamitos, CA, USA, pp. 779–788.

Ruchay, Alexey, Kober, Vitaly, Dorofeev, Konstantin, Kolpakov, Vladimir, Miroshnikov, Sergei, 2020. Accurate body measurement of live cattle using three depth cameras and non-rigid 3-D shape recovery. Comput. Electron. Agric. 179, 105821.

Ruchay, Alexey, Kolpakov, Vladimir, Guo, Hao, Pezzuolo, Andrea, 2024. On-barn cattle facial recognition using deep transfer learning and data augmentation. Comput. Electron. Agric. 225, 109306.

Sharma, Asheesh, Randewich, Lucy, Andrew, William, Hannuna, Sion, Campbell, Neill, Mullan, Siobhan, Dowsey, Andrew W., Smith, Melvyn, Hansen, Mark, Burghardt, Tilo, 2025. Universal bovine identification via depth data and deep metric learning. Comput. Electron. Agric. 229, 109657.

Shorten, Connor, Khoshgoftaar, Taghi M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 1–48.

Sun, Ke, Xiao, Bin, Liu, Dong, Wang, Jingdong, 2019. Deep high-resolution representation learning for human pose estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5686–5696.

Wang, Chien-Yao, Bochkovskiy, Alexey, Liao, Hong-Yuan Mark, 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 7464–7475.

Wang, Zhenyao, Liu, Tonghai, 2022. Two-stage method based on triplet margin loss for pig face recognition. Comput. Electron. Agric. 194, 106737.

Wang, Yaowu, Mücher, Sander, Wang, Wensheng, Kooistra, Lammert, 2024. Automated retrieval of cattle body measurements from unmanned aerial vehicle-based LiDAR point clouds. Comput. Electron. Agric. 227, 109521.

Yang, Guangyuan, Xu, Xingshi, Song, Lei, Zhang, Qianru, Duan, Yuanchao, Song, Huaibo, 2022. Automated measurement of dairy cows body size via 3D point cloud data analysis. Comput. Electron. Agric. 200, 107218.

Zhang, Na, 2023. A study on the impact of face image quality on face recognition in the wild. ArXiv, abs/2307.02679.

Zheng, Zhaohui, Wang, Ping, Ren, Dongwei, Liu, Wei, Ye, Rongguang, Hu, Qinghua, Zuo, Wangmeng, 2021. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Trans. Cybern..

Zhou, Kaiyang, Yang, Yongxin, Cavallaro, Andrea, Xiang, Tao, 2019. Omni-scale feature learning for person re-identification. In: 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 3701–3711.