AgriEngineering

MDPI

Article

DINOv2-Driven Monocular Body Measurement Keypoint Detection for Low-Texture Endangered Binglangjiang Buffalo

Yuhan Xun, Xingchen Ye, Yinuo He, Bo Hu and Fei Xiong *

School of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming 650224, China; xunyuhan10@gmail.com (Y.X.); chen@swfu.edu.cn (X.Y.); dongruihyn@outlook.com (Y.H.); bohu3090@outlook.com (B.H.)

* Correspondence: xiongfei@swfu.edu.cn

Abstract

The Binglangjiang buffalo, the only indigenous river-type buffalo in China, poses significant challenges for automated keypoint detection due to its uniformly black, low-texture coat, poor foreground–background contrast, and scarcity of annotated training samples. To address these challenges, this study constructs a benchmark dataset of 10,834 lateral-view images covering 424 individuals, annotated with 10 body measurement keypoints following standardized buffalo measurement protocols. A keypoint detection pipeline is developed by adapting DINOv2 with a top-down heatmap regression head under a single-view imaging setup, reducing hardware complexity for practical farm deployment. Benchmarking against YOLOv8 series and a standard ViT baseline shows that DINOv2-Base achieves 96.51% mAP, surpassing YOLOv8m by 5.6 percentage points. Compared to standard ViT, DINOv2 demonstrates more stable localization across keypoints under model scaling. Specifically, on the scapular tip (P8), a particularly low-texture region, DINOv2 exhibits only 0.28% mAP fluctuation versus 0.82% for standard ViT, indicating greater robustness to limited training data and low-contrast imaging. Body measurement validation on 20 individuals yields MAPE values of 1.76–5.69% across five measurements, confirming reliable non-contact measurement performance. The dataset and pipeline provide practical support for precision livestock management of endangered breeds.

Keywords: Binglangjiang buffalo; keypoint detection; DINOv2; Vision Transformer; low-texture imaging; benchmark dataset; body measurement; precision livestock farming

Academic Editors: Patrícia Ferreira Ponciano Ferraz, Verónica González Cadavid and Giuseppe Rossi

Received: 24 March 2026
Revised: 10 May 2026
Accepted: 21 May 2026
Published: 1 June 2026

1. Introduction

The Binglangjiang buffalo (*Bubalus bubalis*) is the only river-type buffalo native to China, distributed primarily in the Binglangjiang River Valley of Tengchong City, Yunnan Province [1], and has been formally listed in the *National Inventory of Protected Livestock and Poultry Genetic Resources* [2]. This breed exhibits substantially higher milk yields than domestic swamp-type buffaloes, with peak production reported to exceed 3000 kg [1], and further demonstrates adaptive advantages including tolerance to coarse forage and strong disease resistance, conferring considerable value as a protected genetic resource.

In the context of conservation and selective breeding programs for this breed, morphometric body measurements, including withers height, body length, and chest girth, serve as core indicators for evaluating individual growth, development, and productive performance. However, accurate acquisition of these measurements remains challenging in

practical farming environments. Most existing approaches rely on multi-view systems or controlled conditions, which limits their applicability in real-world scenarios.

Conventional body measurement relies on manual contact-based procedures, which are associated with low throughput, poor inter-operator consistency, and a propensity to induce stress responses in the animal [3,4]. These limitations render large-scale, periodic data collection impractical under field production conditions. With the increasing adoption of precision livestock farming supported by intelligent and data-driven technologies for improving livestock productivity, welfare, and sustainability [5–8], image-based non-contact body measurement estimation has emerged as an active area of research [9–11]. Three-dimensional point cloud-based approaches have also been explored for buffalo body measurement [12]. In image-based pipelines, a critical prerequisite is the accurate localization of anatomical body measurement keypoints on the animal body surface, from which body measurements can be derived through geometric computation.

However, the application of existing keypoint detection methods to the Binglangjiang buffalo presents two domain-specific challenges. First, limited data availability: as an endangered species with a small extant population, the number of labeled images available for model training is substantially lower than in conventional livestock studies, imposing stricter demands on sample efficiency. Second, low target contrast: the uniformly black coat of the Binglangjiang buffalo, combined with the shadowed environment of semi-open barns, results in high pixel-level similarity between the foreground subject and the dark background, a condition under which detection methods relying on local gradient features are likely to exhibit keypoint localization drift. These two factors collectively constitute the core problem context motivating this study.

In response to the aforementioned challenges, this study presents the following two principal contributions.

(1) Construction of a standardized body measurement keypoint dataset for the Binglangjiang buffalo. Lateral-view images of 424 individuals were systematically acquired under diverse lighting conditions in a semi-open barn environment in Tengchong, Yunnan. Following deduplication and quality filtering, more than 10,000 valid samples were retained. A unified annotation protocol comprising 10 body measurement keypoints (including the highest point of the withers and the sternal base point) was established in accordance with veterinary morphometric standards, and fine-grained manual annotation was completed for the entire dataset. This dataset fills the gap of standardized vision data for the Binglangjiang buffalo and serves as a baseline resource for subsequent body measurement research on this breed.

(2) Development and validation of a DINOv2-based keypoint detection pipeline for buffalo body measurement. To overcome the insufficient discriminative feature extraction capability of CNN backbones under low-texture, limited-sample conditions, DINOv2 [13], a Vision Transformer (ViT) originally designed for general-purpose visual representation learning, was adapted into the buffalo body measurement keypoint detection domain and integrated with a top-down heatmap regression head to construct the detection pipeline. The global self-attention mechanism of the ViT architecture enables the model to capture inter-part structural dependencies across the full body skeleton within a single forward pass, compensating for the absence of reliable local texture cues that CNN-based backbones depend upon. With all models trained from scratch under identical conditions, systematic experiments were conducted against both mainstream YOLOv8 series models and a standard ViT baseline. The results demonstrate that DINOv2 maintains stable localization performance on keypoints with indistinct surface morphological features where the standard ViT exhibits notable accuracy degradation, providing empirical evidence for

pipeline selection in keypoint detection tasks involving analogous endangered species under limited-sample conditions.

The entire pipeline operates under a single-view imaging setup, providing a practical and deployable solution for livestock body measurement in farm environments.

2. Related Work

Animal keypoint detection methodology has broadly evolved along three successive paradigms: direct coordinate regression, probabilistic heatmap regression, and Transformer-backbone-based approaches. The heatmap regression paradigm encompasses both top-down [14,15] and bottom-up [16] formulations. Toshev and Szegedy pioneered the application of convolutional neural networks (CNN) to directly regress human joint coordinates, establishing the foundational framework for deep learning-based pose estimation [17]. While groundbreaking, direct coordinate regression is susceptible to spatial information loss and quantization errors. Sun et al. subsequently introduced the soft-argmax operation to improve training convergence [14], though purely regression-based methods remain limited under complex articulation and deformation scenarios. Heatmap regression methods, exemplified by Stacked Hourglass Networks [15] and HRNet [18], addressed these limitations by maintaining high-resolution feature representations throughout the network, establishing long-standing state-of-the-art performance on general pose estimation benchmarks. In the domain of animal pose estimation, the DeepLabCut framework proposed by Mathis et al. demonstrated that CNN backbones pre-trained on ImageNet could achieve high-precision keypoint tracking of laboratory animals through fine-tuning on limited labeled samples [19]; this paradigm was subsequently extended to pose estimation of large livestock species, including cattle [20] and pigs [21]. Building on heatmap regression baselines [22], more recently, Vision Transformer-based methods such as ViTPose [23] have further advanced detection accuracy across multiple pose estimation benchmarks, suggesting that Transformer architectures [24] may offer inherent advantages in modeling global structural dependencies.

3. Materials and Methods

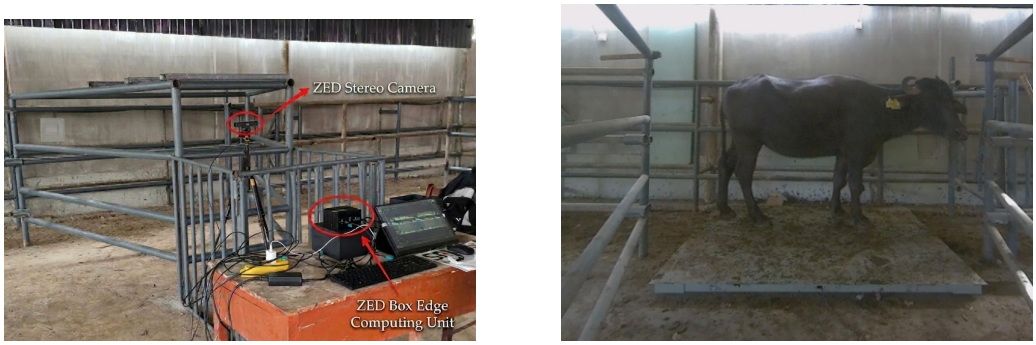
This section describes the complete experimental framework underlying the proposed pipeline. The dataset construction process, including image acquisition and quality filtering, is detailed first, followed by the keypoint annotation protocol and the corresponding body measurement framework. The data partitioning and augmentation strategies are then described, along with the network architecture comprising the DINOv2 feature extraction backbone and the heatmap prediction head. Finally, the hardware and software configuration and the evaluation metrics adopted throughout this study are presented.

3.1. Dataset Construction

Image Acquisition

Experimental images were acquired at the Bafu Le Binglangjiang Buffalo Farm in Hehua Town, Tengchong City, Yunnan Province, China. A ZED 2 stereo camera (Stereolabs, Paris, France) in conjunction with a ZED Box edge computing module was employed for image capture, on-site storage, and data transmission. To facilitate subsequent keypoint annotation, individual animals were sequentially guided into a designated capture zone to obtain standardized lateral-view images under controlled framing conditions. All images were acquired under a single-view imaging setup, with only the left camera stream used for subsequent processing. The resulting dataset encompasses all 424 animals maintained at the facility, comprising 13,423 raw images acquired predominantly at a resolution of

1120 × 800 pixels, with a minor subset recorded at 1120 × 680 pixels. The data acquisition environment and a representative sample image are illustrated in Figure 1.



(**a**) On-site camera deployment (**b**) Standardized lateral-view sample

Figure 1. Data acquisition setup. (**a**) On-site deployment of the ZED 2 stereo camera in the semi-open barn at the Bafu Le Binglangjiang Buffalo Farm, with the camera installation position indicated. (**b**) Representative lateral-view image captured under the standardized acquisition protocol.

To mitigate inter-sample redundancy arising from excessively short acquisition intervals, a minimum temporal gap was enforced between consecutive captures of the same individual. To further enhance data diversity and reduce potential distributional bias at the acquisition stage, samples were deliberately collected across a range of ambient lighting conditions and at multiple time points throughout the day. Since all 424 Binglangjiang buffalo individuals maintained at the facility were included, the dataset covers the available farm population rather than a manually selected subset.

3.2. Keypoint Definition and Body Measurement Framework

Ten body measurement keypoints were defined in accordance with established veterinary conventions and the practical requirements of morphometric body measurement, as illustrated in Figure 2.

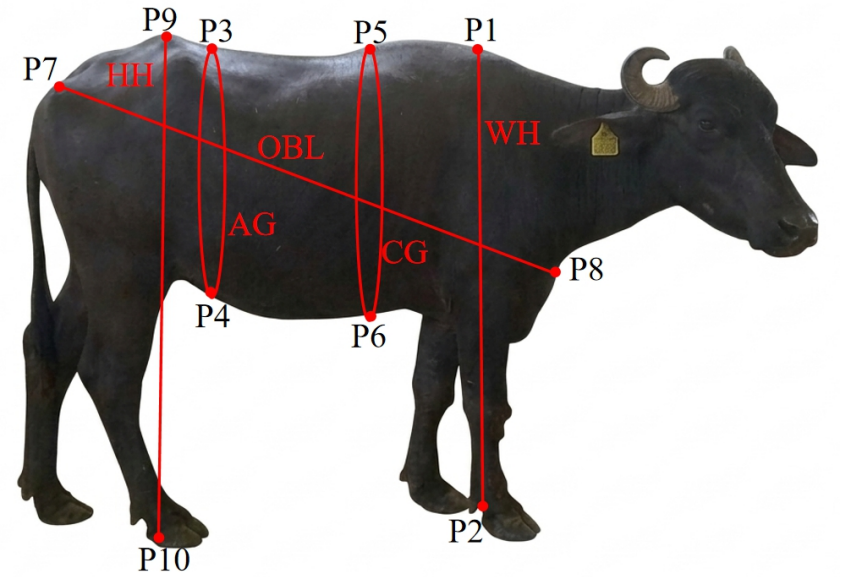


Figure 2. Body measurement keypoint definitions and corresponding body measurement parameters for the Binglangjiang buffalo. Vertical lines indicate height measurements (WH, HH); the diagonal line indicates oblique body length (OBL); elliptical arcs indicate girth measurements (CG, AG).

The ten keypoints are defined as follows: P1, the highest point of the withers; P2, the contact point between the fore-hoof and the ground; P3, the lumbar vertebra point; P4, the lowest point of the ventral midline at the abdominal region; P5, the posterior border point of the withers; P6, the sternal base point; P7, the posterior border point of the ischial tuberosity; P8, the anterior border point of the scapular tip; P9, the highest point of the lumbar-sacral junction; and P10, the contact point between the hind-hoof and the ground.

Based on these keypoints, five principal body measurements are derived: withers height (WH), measured as the vertical distance from P1 to the ground plane; hip height (HH), measured as the vertical distance from P9 to the ground plane; oblique body length (OBL), measured as the straight- line distance from P8 to P7; chest girth (CG), measured as the body circumference at the cross-section defined by P5 and P6; and abdominal girth (AG), measured as the maximum body circumference at the cross-section defined by P3 and P4.

The selection of these five measurement targets was guided by established livestock morphometric practices, but was adapted to the specific objective of monocular trunk body measurement. Complete phenotypic characterization protocols may include a broader set of traits, such as horn, ear, tail, or udder-related traits, in addition to trunk-related body measurements [25,26]. However, automated body-size measurement studies commonly focus on a compact set of trunk-related measurements, such as withers height, body length or oblique body length, chest girth, hip height, and abdominal or paunch girth [3,4,9,10,12,27]. Therefore, the five measurements selected in this study—withers height, hip height, oblique body length, chest girth, and abdominal girth—were chosen because they are directly related to body size, biologically meaningful for farm management, and feasible to estimate from standardized lateral-view images.

Accordingly, the ten keypoints were defined as a compact anatomical landmark set required for these five target measurements, rather than as an exhaustive representation of all possible buffalo phenotypic traits. P1–P2 and P9–P10 provide the anterior and posterior vertical height references, P8–P7 describes the longitudinal trunk dimension, P5–P6 defines the thoracic cross-section for chest girth estimation, and P3–P4 defines the abdominal cross-section for abdominal girth estimation. These keypoint pairs complement each other by covering the vertical, longitudinal, thoracic, abdominal, and posterior body regions required for the proposed monocular body-measurement framework.

3.3. Data Partitioning and Augmentation

Prior to partitioning, the raw collection of 13,423 images underwent a systematic quality filtering procedure to ensure annotation reliability. Images were excluded based on the following criteria: (1) motion blur, caused by excessive animal movement during capture; (2) incomplete body coverage, where a substantial portion of the buffalo's body extended beyond the image boundary; (3) multi-instance overlap, where two or more individuals simultaneously entered the capture zone and occluded each other, precluding unambiguous single- animal annotation; and (4) non-lateral orientation, where the subject presented a frontal or posterior view inconsistent with the standardized lateral- view acquisition protocol. Following this filtering procedure, 10,834 images were retained as the final annotated dataset.

For model development and evaluation, the annotated dataset of 10,834 images was partitioned into a training set (7588 images) and a test set (3246 images) using a 7:3 ratio through a randomized sampling procedure. This partitioning strategy ensures that both subsets encapsulate the full spectrum of environmental conditions—including diverse illumination, various animal poses, and complex background variations—thereby pro-

viding a robust foundation for evaluating the model's generalization across the breed's morphological diversity in real-world farm settings.

To remain compatible with the patch-based architecture of the DINOv2 backbone, all input images were resized and padded to a unified resolution of 518 × 518 pixels. During the training phase, an online data augmentation pipeline was implemented within the MMPose framework to enhance model robustness. This included: (1) Geometric transformations: random horizontal flipping ($p = 0.5$), random rotation within ±40°, and random scaling factors (0.5–1.5×); (2) Half-body transform: applied with a probability of 0.3 to simulate partial occlusions typically encountered in farm environments; and (3) Unbiased Data Processing (UDP) [28]: integrated into the transformation and heatmap generation stages to minimize coordinate quantization errors and refine localization precision. No augmentation was applied to the test set, which was reserved exclusively for final performance evaluation.

3.4. Network Architecture

An encoder–decoder architecture was adopted for keypoint detection, comprising two principal components: a feature extraction backbone and a heatmap prediction head. The overall pipeline is schematically illustrated in Figure 3.

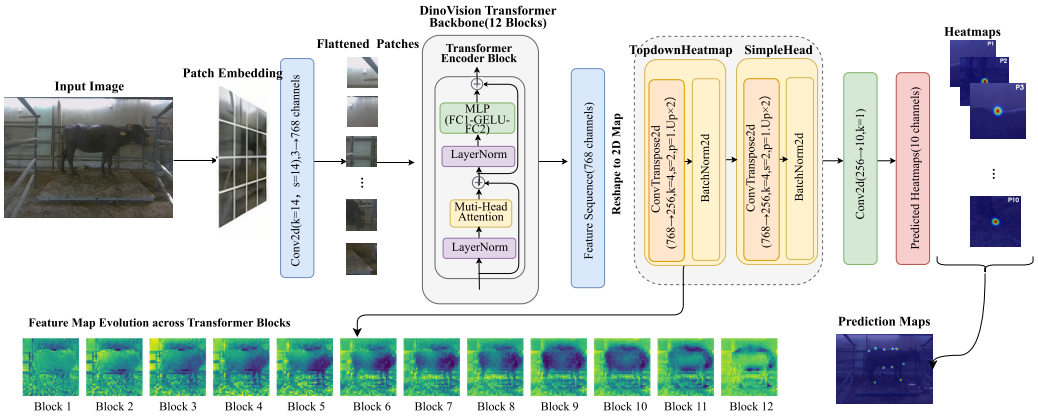


Figure 3. Schematic overview of the proposed encoder–decoder keypoint detection pipeline.

3.4.1. Feature Extraction Backbone

DINOv2-Base [13,29] was employed as the feature extraction backbone. In contrast to convolutional neural networks, whose feature extraction is inherently constrained to local receptive fields, the self-attention mechanism [30] of the ViT models pairwise dependencies between arbitrary spatial positions within a single forward pass. This global receptive field property is of particular relevance in low-texture scenarios such as the Binglangjiang buffalo: when local regions lack identifiable surface features due to the uniformly dark coat, the model can draw upon global morphological and structural context across the entire body skeleton to support localization, potentially conferring a fundamental advantage over locally-operated CNN architectures.

Input images were uniformly resized to 518 × 518 pixels prior to processing. The patch embedding layer partitioned each image into non-overlapping 14 × 14-pixel patches, which were subsequently flattened and passed through a 12-layer Transformer encoder. The resulting output feature sequence was rearranged into a two-dimensional spatial feature tensor of shape 37 × 37 × 768, which serves as the input to the prediction head.

3.4.2. Heatmap Prediction Head

A lightweight top-down heatmap regression head [22] was appended to the backbone to produce per-keypoint localisation maps. The head consists of a deconvolutional upsam-

pling module followed by a 1×1 convolutional layer that projects the feature tensor onto $K = 10$ channels, each corresponding to one body measurement keypoint. Each output channel is interpreted as a Gaussian probability heatmap, and the predicted keypoint coordinate is obtained as the spatial location of the maximum activation. The mean squared error (MSE) between the predicted heatmaps and the ground-truth Gaussian targets was adopted as the training objective.

3.4.3. Geometric Computation of Body Measurements

Body measurements were derived from detected keypoint coordinates in conjunction with the corresponding depth values through standard perspective projection inversion. Let (u_i, v_i) denote the pixel coordinate of keypoint P_i and Z_i the associated depth value in metres. The metric three-dimensional position of each keypoint was reconstructed as:

$$X_i = \frac{(u_i - c_x)\, Z_i}{f_x}, \quad Y_i = \frac{(v_i - c_y)\, Z_i}{f_y} \tag{1}$$

where f_x and f_y denote the horizontal and vertical focal lengths of the left camera in pixels, and (c_x, c_y) denotes the principal point, all obtained from the factory calibration.

WH was computed as the Euclidean distance between the 3D coordinates of P_1 and P_2, and HH was computed as the Euclidean distance between the 3D coordinates of P_9 and P_{10}:

$$H_{\mathrm{WH}} = \sqrt{(X_{P_1} - X_{P_2})^2 + (Y_{P_1} - Y_{P_2})^2 + (Z_{P_1} - Z_{P_2})^2} \tag{2}$$

$$H_{\mathrm{HH}} = \sqrt{(X_{P_9} - X_{P_{10}})^2 + (Y_{P_9} - Y_{P_{10}})^2 + (Z_{P_9} - Z_{P_{10}})^2} \tag{3}$$

OBL was computed as the Euclidean distance between the 3D coordinates of P_7 and P_8:

$$L_{\mathrm{OBL}} = \sqrt{(X_{P_7} - X_{P_8})^2 + (Y_{P_7} - Y_{P_8})^2 + (Z_{P_7} - Z_{P_8})^2} \tag{4}$$

CG and abdominal girth (AG) were estimated via path integration along the visible lateral body contour between keypoints P5–P6 and P3–P4, respectively. For each girth measurement, Ns = 100 points were sampled at uniform intervals along the line segment connecting the two keypoints in image space, and the corresponding 3D coordinates were reconstructed via depth back-projection.

Such contour-based geometric estimation strategies are conceptually related to classical model fitting and shape reconstruction methods widely used in computer vision and geometric analysis [31,32]. The visible arc length was then computed as the cumulative sum of consecutive 3D segment lengths and multiplied by a factor of 2 to approximate the full body circumference.

$$C_{\mathrm{raw}} = 2 \sum_{j=1}^{N_s - 1} \left\| \mathbf{p}_{j+1} - \mathbf{p}_j \right\|_2 \tag{5}$$

where $\mathbf{p}_j \in \mathbb{R}^3$ denotes the reconstructed 3D position of the j-th sampled point.

To compensate for the systematic underestimation introduced by single-view path integration, linear calibration coefficients α_k were applied:

$$\hat{m}_k = \alpha_k \cdot m_k^{\mathrm{raw}} \tag{6}$$

where m_k^{raw} is the raw predicted value and $\hat{m}_k$ is the calibrated measurement. Coefficients were determined using leave-one-out cross-validation (LOOCV) across the 20 validation individuals, yielding $\alpha_{\mathrm{WH}} = 1.049$, $\alpha_{\mathrm{HH}} = 1.017$, $\alpha_{\mathrm{OBL}} = 1.070$, $\alpha_{\mathrm{CG}} = 1.214$, and $\alpha_{\mathrm{AG}} = 1.320$.

The body measurement evaluation was conducted on a separate set of individuals with independently collected ground-truth measurements, where individual identities were explicitly recorded. This evaluation protocol is independent of the keypoint detection dataset and provides additional support for the reliability of the proposed method.

3.5. Experimental Setup and Evaluation Metrics

3.5.1. Hardware and Software Configuration

All model training and inference were performed on a workstation equipped with two Intel Xeon Silver 4210R CPUs (total 20 cores, 40 threads) and an NVIDIA GeForce RTX 3090 GPU (24 GB VRAM). The software environment was built upon AlmaLinux OS 9.7, utilizing CUDA 11.3, PyTorch 1.11.0, and Python 3.9.

The AdamW optimizer was employed with an initial learning rate of 5×10^{-4} and a weight decay of 0.1. To optimize the Vision Transformer backbone, a layer-wise learning rate decay of 0.75 was implemented. The learning rate was scheduled using a step decay policy. A linear warmup strategy was applied during the first 500 iterations with a warmup ratio of 0.001. Batch sizes were set to 8 per GPU, determined by the available VRAM capacity. All models were trained and evaluated under identical hardware conditions, with performance metrics computed uniformly via the COCO API.

3.5.2. Evaluation Metrics

The primary evaluation metric adopted in this study is mean Average Precision (mAP) computed on the basis of Object Keypoint Similarity (OKS) [33], which is defined as:

$$OKS = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\,\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{7}$$

where d_i denotes the Euclidean distance between the predicted keypoint and its ground-truth annotation, s is the object scale factor, k_i is the per-keypoint normalization constant [33], and $\delta(\cdot)$ is the visibility indicator function. The default k_i values from the COCO human pose estimation benchmark were adopted without recalibration for buffalo-specific keypoints, which may limit discriminability at higher OKS thresholds; the validity of this approximation warrants empirical verification in subsequent work. The mAP was computed as the mean AP over OKS thresholds ranging from 0.5 to 0.95 in increments of 0.05. AP@0.5 and AP@0.75 are additionally reported as supplementary metrics to provide a more complete characterization of localization precision.

4. Results

This section presents the experimental results of the proposed pipeline across four evaluation dimensions. Training convergence behaviour for all model configurations is first reported, followed by a comparative evaluation of DINOv2 against the YOLOv8-pose model family. A controlled backbone architecture ablation study is then conducted to isolate the effects of architectural design and parameter scale on localization performance. Per-keypoint detection accuracy is subsequently analysed to characterise model behaviour across anatomically distinct regions, with representative failure cases discussed. Finally, body measurement accuracy is validated on 20 individuals via leave-one-out cross-validation.

4.1. Training Performance

The normalized training loss curves for all four model configurations are presented in Figure 4. Loss values were normalized by each model's initial loss to enable cross-architecture comparison on a unified scale. The training of the DINOv2-Large model

involved a checkpoint-based resumption at epoch 50, marked by the vertical dotted line in the figure, to ensure stable convergence of the high-parameter architecture.

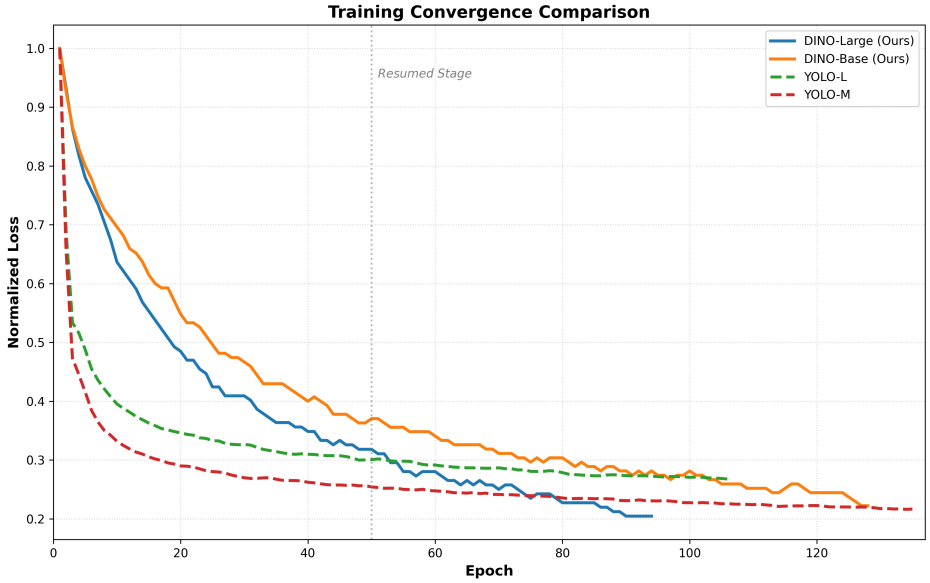


Figure 4. Normalized training loss curves for DINOv2-Base, DINOv2-Large, YOLOv8m-pose, and YOLOv8l-pose over the full training schedule. Loss values are normalized by each model's initial loss to facilitate cross-architecture comparison. The vertical dotted line indicates the checkpoint resumption point for DINOv2-Large at epoch 50.

All models exhibited monotonic convergence without divergence, confirming stable optimization throughout training. The two YOLO variants converged more rapidly in the early epochs, attributable to their lightweight YOLOv8 Backbone. In contrast, the DINOv2 models converged more gradually, consistent with the larger parameter scale and smaller batch size of these architectures.

Notably, DINOv2-Large achieved a lower final training loss than DINOv2-Base (approximately 0.20 vs. 0.25 on the normalized scale), yet obtained an inferior mAP of 96.23% compared to 96.51% for DINOv2-Base on the held-out test set. This divergence between training loss and evaluation performance provides direct empirical evidence for the overfitting hypothesis discussed in Section 4.3: the larger parameter capacity of DINOv2-Large exceeds what the current dataset scale can effectively regularize, resulting in stronger fitting to training distribution at the cost of generalization.

4.2. Comparative Experiment

To evaluate the effectiveness of the DINOv2 backbone on the constructed dataset, a systematic comparison was conducted against the YOLOv8-pose model family. All models were trained and assessed under identical hardware and software conditions, with evaluation performed uniformly using the COCO API to ensure metric comparability. To further assess the stability of the reported results, 95% confidence intervals (CIs) of mAP were estimated by bootstrap resampling on the test set. Results are summarized in Table 1.

DINOv2-Base achieved the highest mAP of 96.51%, with a 95% bootstrap CI of [95.67%, 96.76%]. This result exceeded YOLOv8m-pose (90.95%, 95% CI: [90.03%, 91.80%]) and YOLOv8l-pose (89.95%, 95% CI: [88.88%, 90.96%]) by approximately 5.6 and 6.6 percentage points, respectively. The non-overlapping confidence intervals between DINOv2-Base and the YOLOv8-pose variants indicate that the observed performance advantage is stable under bootstrap resampling.

Table 1. Performance comparison of different models on the Binglangjiang buffalo keypoint dataset. The 95% confidence intervals (CIs) of mAP were estimated by bootstrap resampling on the test set.

Model	Backbone	Params (M)	FPS	mAP (%)	95% CI (%)
YOLOv8m-pose	YOLOv8 Backbone	26.4	327.2	90.95	[90.03, 91.80]
YOLOv8l-pose	YOLOv8 Backbone	43.7	215.5	89.95	[88.88, 90.96]
DINOv2-Base	ViT-Base	90.0	47.00	96.51	[95.67, 96.76]
DINOv2-Large	ViT-Large	308.5	16.82	96.23	[95.41, 96.43]

This performance gap is consistent with the architectural difference in feature extraction strategy: the CNN-based backbone of YOLOv8 operates through spatially local convolutional kernels, whereas the ViT [24] processes all spatial positions jointly within each forward pass. Whether this architectural distinction is the primary driver of the observed accuracy difference under low-texture conditions warrants further controlled investigation on independently collected datasets to fully isolate architectural effects from data-specific influences; the present results nonetheless demonstrate that the ViT-based pipeline achieves superior keypoint localisation on this dataset.

With respect to inference throughput, YOLOv8m achieved 327.2 FPS, substantially exceeding the 47.00 FPS of DINOv2-Base. For non-real-time batch body measurement applications, 47.00 FPS is considered adequate; however, real-time deployment on resource-constrained embedded devices remains challenging, representing a principal limitation of the proposed pipeline.

4.3. Backbone Architecture Ablation Study

To investigate the intrinsic effects of model architecture and parameter scale on keypoint detection performance, a controlled ablation study was conducted across backbone configurations. Bootstrap-based 95% confidence intervals were also reported for mAP to provide a more stable numerical comparison across backbone variants. Results are presented in Table 2.

Table 2. Ablation study results across backbone architectures and parameter scales. The 95% confidence intervals (CIs) of mAP were estimated by bootstrap resampling on the test set.

Model	Params	FPS	AP@0.5	AP@0.75	mAP (95% CI)
ViT-Base	90.0 M	88.31	98.02	98.02	96.53 [95.72, 96.82]
ViT-Large	308.5 M	47.35	98.02	98.02	95.71 [94.82, 95.91]
DINOv2-Base	90.8 M	47.00	98.02	98.02	96.51 [95.67, 96.76]
DINOv2-Large	309.6 M	16.82	98.02	98.02	96.23 [95.41, 96.43]

Note: The identical AP@0.5 and AP@0.75 values across all configurations are primarily attributable to the use of default COCO k_i normalization coefficients, whose broad tolerance thresholds limit metric discriminability at higher OKS levels.

At the Base scale, ViT-Base and DINOv2-Base exhibited highly convergent mAP values of 96.53% and 96.51%, with overlapping 95% confidence intervals of [95.72%, 96.82%] and [95.67%, 96.76%], respectively. This indicates that their overall mAP performance is comparable at the aggregate level. However, aggregate metric similarity does not necessarily imply equivalent behaviour under challenging localization conditions. Per-keypoint analysis via radar chart (Figure 5) reveals a meaningful performance divergence between the two architectures at P8 (anterior border of the scapular tip), a particularly challenging low-texture keypoint.

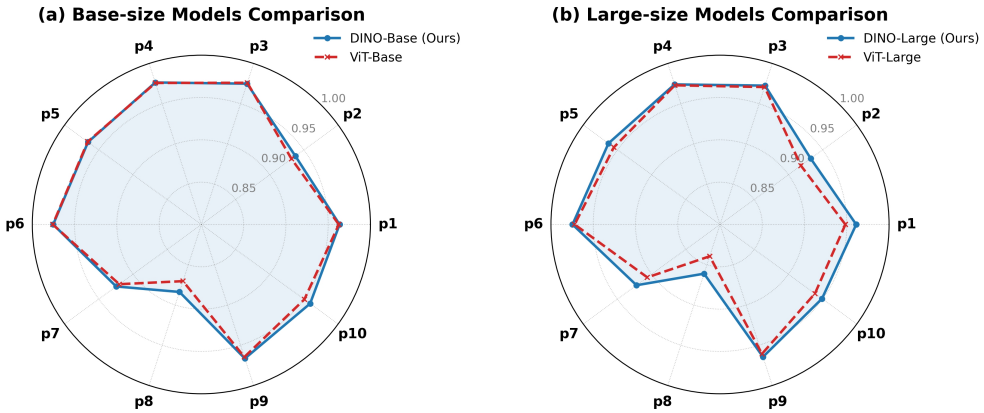


Figure 5. Radar chart comparing per-keypoint AP across the 10 body measurement keypoints for all backbone configurations. The DINO architecture demonstrates comparatively stronger localization performance at challenging low-texture keypoints such as P8.

Due to the uniformly dark coat of the Binglangjiang buffalo, the shoulder region encompassing P8 exhibits minimal local contrast and lacks geometric discontinuities, making it inherently susceptible to localization drift. This architectural difference is further amplified when scaling from Base to Large configurations: the ViT backbone exhibits a substantial accuracy decline at P8 in the Large variant (dropping to 83.95%), whereas the DINO backbone maintains performance above 86.11% under the same conditions. The capacity to sustain high localization accuracy in regions with limited discriminative texture may be attributed to the global receptive field of the ViT architecture, which inherently encodes inter-part structural dependencies across the full body skeleton within each forward pass, rather than relying on local texture gradients as in convolutional operations, although this interpretation remains inferential and requires further validation. The more stable scaling behaviour observed in the DINO variant further suggests that its architectural design contributes to more consistent feature representations under the current dataset scale.

Furthermore, the two architectures differ markedly in their stability under parameter scaling. Increasing the parameter count from approximately 90 M to over 300 M yields a statistically meaningful mAP decline of 0.82% in the ViT architecture, suggesting a tendency toward overfitting at the current dataset scale of approximately 10,000 images. In contrast, the DINO architecture exhibits only a 0.28% fluctuation under equivalent scaling, indicating more robust feature generalization. Collectively, while DINO-Base is moderately slower in inference throughput than ViT-Base, its superior robustness at challenging keypoints and greater stability under model scaling render it the more suitable choice for practical deployment in digitalized livestock breeding applications.

4.4. Per-Keypoint Detection Accuracy Analysis

Individual AP values for DINOv2-Base across all 10 keypoints are reported in Table 3.

Per-keypoint AP values ranged from 88.38% to 97.64%, and the observed variation is consistent with the surface morphological characteristics and imaging conditions associated with each keypoint. P1, P4, P6, and P9 achieved AP values exceeding 96%, as they correspond to well-defined bony prominences or morphological extrema of the body contour.

The two lowest-performing keypoints were P8 (88.38%) and P7 (92.46%). P8 is located within the muscle-covered region anterior to the scapula where the surface transitions smoothly without discernible features, rendering it the most challenging keypoint. P7 corresponds to the posterior border of the ischial tuberosity, a gradually transitioning bony prominence without a well-defined morphological extremum, whose surface is

further obscured by overlying gluteal musculature, resulting in a smooth contour region with limited local visual anchors for precise localization. Nevertheless, both keypoints maintained AP values above 88%, suggesting that the model retains an acceptable baseline localization capability.

Table 3. Per-keypoint AP values for the DINOv2-Base model.

ID	Keypoint	AP (%)
P1	The highest point of the withers	96.35
P2	The contact point between the fore-hoof and the ground	93.75
P3	The lumbar vertebra point	97.49
P4	The lowest point of the ventral midline at the abdominal region	97.64
P5	The posterior border point of the withers	96.57
P6	The sternal base point	97.53
P7	The posterior border point of the ischial tuberosity	92.46
P8	The anterior border point of the scapular tip	88.38
P9	The highest point of the lumbar-sacral junction	96.65
P10	The contact point between the hind-hoof and the ground	95.90
Overall	**mAP**	**96.51**

4.5. Failure Case Analysis

A subset of test samples exhibited detection errors, primarily concentrated in two scenarios: multi-instance interference, where a second individual appeared at the image boundary causing prediction bias towards the interfering subject; and limb misidentification, where the near and far hindlimbs overlapped in the 2D lateral view, causing keypoints to be localized on the incorrect limb (Figure 6).

Multi-instance interference occurs when individuals appear at image boundaries, causing predictions to bias toward interfering subjects. Limb misidentification arises from the overlap of near and far hindlimbs in the 2D lateral projection, which deprives the model of the depth cues necessary to distinguish between them. Incorporating depth information or multi-view fusion may mitigate these errors in future work.

(**a**) Multi-instance interference (**b**) Limb misidentification

Figure 6. Representative failure cases observed in the test set. The keypoints (P1–P10) are shown as markers.

4.6. Body Measurement Results

The body measurement results obtained via leave-one-out cross-validation across 20 individuals (which were independent of the keypoint detection dataset and explicitly recorded with ground-truth measurements) are summarized in Table 4 and Figure 7. To further characterize the stability of the measurement results, 95% confidence intervals (CIs)

of MAPE were also reported for each body measurement parameter. The MAPE values ranged from 1.76% to 5.69% across all five body measurements.

Table 4. Body measurement accuracy of the proposed framework evaluated on 20 Binglangjiang buffalo individuals using leave-one-out cross-validation. The 95% confidence intervals (CIs) are reported for MAPE.

Parameter	MAE (cm)	MAPE (%)	95% CI of MAPE (%)	Max Error (cm)
WH	2.29	1.76	[1.24, 2.29]	5.85
HH	2.54	2.01	[1.55, 2.49]	5.17
OBL	7.49	5.30	[3.70, 7.18]	21.73
CG	8.62	4.44	[3.25, 5.82]	25.10
AG	12.98	5.69	[4.27, 7.31]	31.77

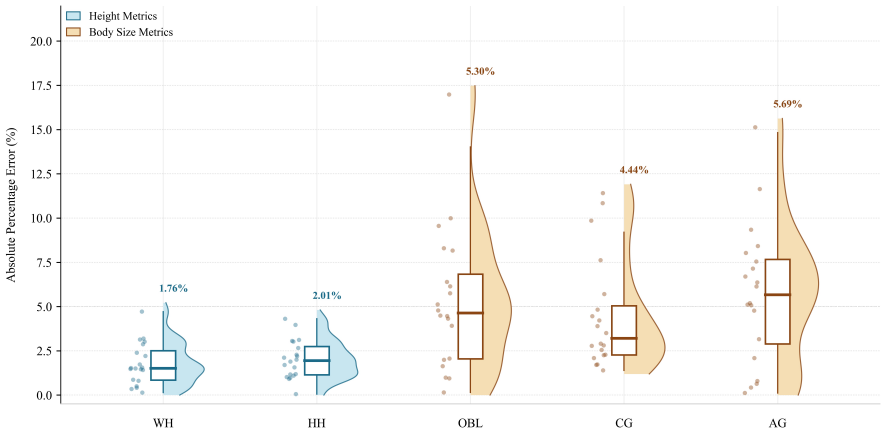


Figure 7. Raincloud plot illustrating the absolute error distribution for the five body measurement parameters across the 20 test individuals. The plot combines a half-violin plot for probability density, a boxplot for statistical summaries, and jittered points for individual error variance.

Height-based measurements achieved the highest accuracy, with WH reaching 1.76% MAPE (95% CI: [1.24%, 2.29%]) and HH reaching 2.01% MAPE (95% CI: [1.55%, 2.49%]). These results outperform the single-view keypoint-based method of Yang et al. [27], which reported 6.7% and 4.1% under comparable acquisition conditions.

Girth measurements showed higher errors, with CG and AG yielding MAPE values of 4.44% (95% CI: [3.25%, 5.82%]) and 5.69% (95% CI: [4.27%, 7.31%]), respectively. This is attributable to two compounding factors: the uniformly black coat impairing stereo matching quality in high-curvature body regions, and the single lateral viewpoint constraining girth estimation to the visible contour, which introduces systematic geometric uncertainty in circumference estimation.

Oblique body length also showed a relatively higher MAPE of 5.30% (95% CI: [3.70%, 7.18%]). This may be related to the simultaneous localization of P7 and P8, where P8 achieved the lowest keypoint AP (88.38%), amplifying localization errors in the resulting distance calculation. Despite these limitations, all five body measurements remained within practically acceptable error ranges, and the reported confidence intervals further support the stability of the proposed non-contact body measurement framework under real farm conditions.

5. Discussion

The DINOv2-Base model adopted in this study achieved an mAP of 96.51% on the standard test set. However, in real-world farming environments, uncontrollable factors such as image noise, illumination variation, and background complexity are inevitable

and can significantly affect model performance. Previous studies have reported that environmental variability can affect the robustness of vision-based livestock measurement systems [34]. Therefore, this section evaluates and compares the detection robustness of DINOv2-Base and YOLOv8m under such interference conditions through synthetic perturbation experiments.

5.1. Robustness to Noise Interference

The assessment of model robustness under varying noise conditions is crucial for transitioning automated measurement systems from controlled experimental settings to practical agricultural deployment [35]. While recent one-stage frameworks for cow body measurement have demonstrated high precision in clean environments [27], their stability against environmental perturbations, such as airborne dust or sensor noise, remains a significant challenge. To evaluate this, Gaussian noise, salt-and-pepper noise, and speckle noise were separately superimposed on the original test set images, with noise intensity ranging over $0 \leq \sigma^2 \leq 0.125$. The evaluation metric used was mAP.

As shown in Figure 8, the two models exhibit markedly different sensitivities to noise. Under speckle noise, DINOv2-Base maintains mAP above 0.96 throughout the entire test range, demonstrating stable performance. Under Gaussian noise, when $\sigma^2 = 0.1$, YOLOv8m's mAP drops from 0.8435 to 0.3233, corresponding to an absolute decrease of 0.5202, while DINOv2-Base remains at 0.9358. Under salt-and-pepper noise, YOLOv8m's mAP falls to near zero at $\sigma^2 = 0.05$, whereas DINOv2-Base maintains 0.7543 under the same condition.

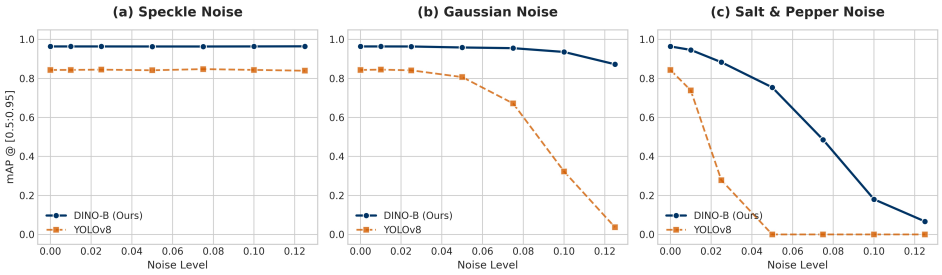


Figure 8. mAP as a function of noise intensity for DINOv2-Base and YOLOv8m under three noise types.

These differences are consistent with the distinct feature extraction paradigms of the two architectures. The CNN backbone in YOLOv8m relies on local convolutional kernels; as noise intensity increases, the signal within local receptive fields is progressively corrupted, which is consistent with observations that purely convolutional one-stage networks can be sensitive to local spatial corruption [27]. DINOv2-Base, by contrast, processes all spatial positions jointly within each forward pass, which may contribute to its more gradual performance degradation under the same conditions. The underlying mechanism warrants further investigation; the present results nonetheless demonstrate that the ViT-based pipeline exhibits substantially greater robustness to synthetic noise perturbations on this dataset.

5.2. Adaptability to Illumination Variation

To assess the adaptability of the proposed model to illumination fluctuations, Gamma transformation was applied to simulate different lighting conditions. Five representative γ values were selected, namely 0.5, 0.8, 1.0, 1.4, and 2.0, where $\gamma < 1$ corresponds to low-illumination scenes and $\gamma > 1$ corresponds to brighter or overexposed scenes. In addition to qualitative visualization, quantitative evaluation was conducted on a fixed subset of 100 test images using mAP as the evaluation metric.

As shown in Figure 9, DINOv2-Base maintains stable keypoint localization under mild to moderate illumination variation. Quantitatively, the model achieves mAP values of 0.935, 0.965, and 0.909 at $\gamma = 0.8$, $\gamma = 1.0$, and $\gamma = 1.4$, respectively. These results indicate that moderate illumination changes do not substantially affect the overall localization accuracy, and most anatomical keypoints can still be predicted reliably when the body contour and major structural cues remain visible.

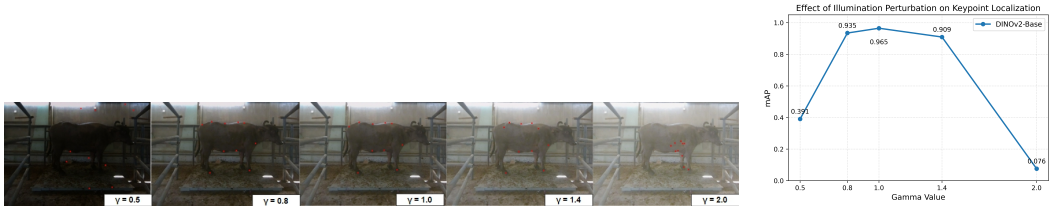


Figure 9. Visualization and quantitative evaluation of DINOv2-Base keypoint predictions under different Gamma illumination transformations. The left panel shows representative prediction results with increasing γ from left to right, while the right panel shows the corresponding mAP variation under selected Gamma values.

However, model performance decreases substantially under extreme illumination changes. Under severe low illumination ($\gamma = 0.5$), the mAP drops to 0.391. This is mainly because the contrast between the buffalo's black coat and the shadowed background becomes extremely low, making body boundaries and weak-texture anatomical landmarks difficult to distinguish. Under severe overexposure ($\gamma = 2.0$), the mAP further decreases to 0.076, suggesting that highlight saturation and contour detail loss can severely disturb keypoint localization. In such cases, the visual cues required for identifying body edges and local anatomical positions are weakened, leading to a higher probability of keypoint drift.

Overall, these results demonstrate that DINOv2-Base is robust to moderate illumination variation, especially within $\gamma = 0.8$ to $\gamma = 1.4$, but remains sensitive to extreme low-light and overexposed conditions. This finding is consistent with the visual results and highlights the importance of maintaining reasonable illumination during practical image acquisition in farm environments.

6. Conclusions

This study addressed the dual challenges of limited sample availability and low-contrast targets inherent to body measurement keypoint detection of the Binglangjiang buffalo. By focusing on a high-performance vision backbone, the research achieved precise localization and reliable body measurement under practical farm conditions. The principal findings are summarized as follows.

First, at the level of data foundation, a standardized body measurement keypoint dataset for the Binglangjiang buffalo was constructed from scratch, filling the gap in standardized vision data for this breed. More than 10,000 lateral-view images were curated with a 10-keypoint annotation protocol, providing a baseline resource for automated phenotypic research.

Second, the proposed pipeline demonstrates that transferring DINOv2 into the livestock domain effectively overcomes the lack of discriminative texture cues on the buffalo's surface. DINOv2-Base achieved a localization accuracy of 96.51% mAP, significantly outperforming YOLOv8 models and demonstrating more robust localization stability than the standard ViT baseline, suggesting that its architectural priors confer inherent advantages under conditions where local features are unreliable.

Third, experiments indicated that DINOv2-Base (90.0 M parameters) provides an optimal balance between accuracy and inference throughput (47.00 FPS). This efficiency,

coupled with the reliance on a straightforward single-view configuration, significantly reduces hardware complexity and lowers the barrier for large-scale system integration in existing farm infrastructures.

Fourth, body measurement validation confirmed that the detected keypoints provide a reliable geometric foundation for non-contact measurement, with MAPE values as low as 1.76% for height-based metrics. While the inherent geometric constraints of single-view acquisition affect girth estimation accuracy, the overall performance remains robust enough for routine digital management.

Fifth, the five body measurements produced by the pipeline directly support routine farm management decisions: withers height and hip height track individual growth against breed-standard curves, chest girth provides a proxy for body weight estimation, abdominal girth assists in nutritional and pregnancy monitoring, and oblique body length informs conformation-based selection in breeding programs. Future work will explore coupling periodic measurement outputs with individual animal records to build a structured decision-support interface for farm operators.

Despite these contributions, this study has limitations that suggest directions for future work. First, the default k_i normalization coefficients from the COCO benchmark were adopted; future research should establish species-specific k_i values to yield more discriminative evaluation results for buffaloes. Second, although the pipeline is efficient, further optimization through model quantization or knowledge distillation [36] will be explored to enhance real-time performance on resource-constrained edge devices. Third, the localization uncertainty at low-texture keypoints such as P8 reflects the inherent ambiguity in anatomical landmark definition where surface features are absent. Future work may explore fuzzy inference approaches, such as intuitionistic fuzzy pooling (INT-FUP), to explicitly model this positional uncertainty and provide confidence-aware outputs for downstream decision support [37]. Fourth, although the present dataset covers a large number of Binglangjiang buffalo individuals under practical farm conditions, the current evaluation mainly reflects intra-breed and intra-farm performance. The proposed DINOv2-based pipeline may have potential applicability to other large-livestock body measurement scenarios because it captures global morphological and structural dependencies rather than relying solely on local texture cues. However, differences in breed morphology, camera viewpoint, illumination, background complexity, and keypoint annotation protocols may affect cross-domain performance. Future work will further investigate cross-breed and cross-farm validation using independently collected or public livestock datasets under unified keypoint definitions.

7. Patents

The work reported in this manuscript has resulted in the following patent application: "A Method, Apparatus, Electronic Device, and Storage Medium for Body Measurement of Buffalo Based on a Monocular Camera," Chinese Patent Application No. CN202511641226.4, filed on 11 November 2025 (Publication No. CN121120754A, published pending grant).

Author Contributions: Conceptualization, F.X.; methodology, Y.X.; software, Y.X.; validation, Y.X. and X.Y.; formal analysis, Y.X.; investigation, X.Y., Y.H. and B.H.; resources, F.X.; data curation, Y.X., X.Y., Y.H. and B.H.; writing—original draft preparation, Y.X.; writing—review and editing, Y.X.; visualization, Y.X.; project administration, Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Scientific Research Fund of Yunnan Provincial Department of Education (Grant No. 2024J0684) and the Industrial Innovation Program of the Yunnan Talent Support Plan (Grant No. XDYC-CYCX-2024-0021). The APC was funded by the authors.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to the non-invasive nature of the data collection process, which involved only optical observation of Binglangjiang buffalo in their natural farming environment without any physical intervention or disruption to their normal activities.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the findings of this study, including high-resolution SVO2 video streams and annotated buffalo keypoint datasets, are stored in the internal repository of the School of Big Data and Intelligent Engineering, Southwest Forestry University. Due to the significant file size of stereoscopic recordings and the conservation status of the Binglangjiang buffalo, these data are not publicly available. However, processed datasets or specific sequences will be available from the corresponding author upon reasonable request for non-commercial research purposes.

Acknowledgments: The authors would like to thank the School of Big Data and Intelligent Engineering for providing the experimental platform and the NVIDIA RTX 3090 GPU resources. Special thanks are also extended to the owners of the Binglangjiang buffalo farm for their assistance in data collection.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Qu, Z.; Li, D.; Miao, Y.; Shen, X.; Yin, Y.; Ai, Y. Investigation and evaluation on germplasm resources of Binglangjiang buffalo. *J. Yunnan Agric. Univ.* **2008**, *23*, 265–269. (In Chinese) [CrossRef]
2. Yu, X. "Binglangjiang buffalo" was included in the national inventory of protected livestock and poultry genetic resources. *Yunnan Agric.* **2014**, *10*, 74. (In Chinese)
3. Ma, W.; Qi, X.; Sun, Y.; Gao, R.; Ding, L.; Wang, R.; Peng, C.; Zhang, J.; Wu, J.; Xu, Z.; et al. Computer Vision-Based Measurement Techniques for Livestock Body Dimension and Weight: A Review. *Agriculture* **2024**, *14*, 306. [CrossRef]
4. Ma, W.; Sun, Y.; Qi, X.; Xue, X.; Chang, K.; Xu, Z.; Li, M.; Wang, R.; Meng, R.; Li, Q. Computer-Vision-Based Sensing Technologies for Livestock Body Dimension Measurement: A Survey. *Sensors* **2024**, *24*, 1504. [CrossRef]
5. Berckmans, D.; Guarino, M. From the Editors: Precision livestock farming for the global livestock sector. *Anim. Front.* **2017**, *7*, 4–5. [CrossRef]
6. Vlaicu, P.A.; Gras, M.A.; Untea, A.E.; Lefter, N.A.; Rotar, M.C. Advancing Livestock Technology: Intelligent Systemization for Enhanced Productivity, Welfare, and Sustainability. *AgriEngineering* **2024**, *6*, 1479–1496. [CrossRef]
7. Bist, R.B.; Wang, D.; Chai, L.; Xiong, Y. Precision Farming Technologies for Monitoring Livestock and Poultry. *AgriEngineering* **2026**, *8*, 64. [CrossRef]
8. Chen, C.; Zhu, W.; Norton, T. Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning. *Comput. Electron. Agric.* **2021**, *187*, 106255. [CrossRef]
9. Du, A.; Guo, H.; Lu, J.; Su, Y.; Ma, Q.; Ruchay, A.; Marinello, F.; Pezzuolo, A. Automatic livestock body measurement based on keypoint detection with multiple depth cameras. *Comput. Electron. Agric.* **2022**, *198*, 107059. [CrossRef]
10. Bahlo, C.; Dahlhaus, P. Livestock data—Is it there and is it FAIR? A systematic review of livestock farming datasets in Australia. *Comput. Electron. Agric.* **2021**, *188*, 106365. [CrossRef]
11. Sun, Y.; Huo, P.; Wang, Y.; Cui, Z.; Li, Y.; Dai, B.; Li, R.; Zhang, Y. Automatic monitoring system for individual dairy cows based on a deep learning framework that provides identification via body parts and estimation of body condition score. *J. Dairy Sci.* **2019**, *102*, 10140–10151. [CrossRef] [PubMed]
12. Li, J.; Li, Q.; Ma, W.; Xue, X.; Zhao, C.; Tulpan, D.; Yang, S.X. Key Region Extraction and Body Dimension Measurement of Beef Cattle Using 3D Point Clouds. *Agriculture* **2022**, *12*, 1012. [CrossRef]
13. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2024**, arXiv:2304.07193. [CrossRef]
14. Sun, X.; Xiao, B.; Wei, F.; Liang, S.; Wei, Y. Integral Human Pose Regression. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 536–553.
15. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In *Proceedings of the Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 483–499.

16. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
17. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014.
18. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3349–3364. [CrossRef]
19. Mathis, A.; Mamidanna, P.; Cury, K.M.; Abe, T.; Murthy, V.N.; Mathis, M.W.; Bethge, M. DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **2018**, *21*, 1281–1289. [CrossRef]
20. Li, X.; Cai, C.; Zhang, R.; Ju, L.; He, J. Deep cascaded convolutional models for cattle pose estimation. *Comput. Electron. Agric.* **2019**, *164*, 104885. [CrossRef]
21. Riekert, M.; Klein, A.; Adrion, F.; Hoffmann, C.; Gallmann, E. Automatically detecting pig position and posture by 2D camera imaging and deep learning. *Comput. Electron. Agric.* **2020**, *174*, 105391. [CrossRef]
22. Xiao, B.; Wu, H.; Wei, Y. Simple Baselines for Human Pose Estimation and Tracking. In *Proceedings of the Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer: Cham, Switzerland, 2018; pp. 472–487.
23. Xu, Y.; Zhang, J.; Zhang, Q.; Tao, D. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Proceedings of the Advances in Neural Information Processing Systems*; Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 38571–38584.
24. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
25. FAO. *Phenotypic Characterization of Animal Genetic Resources*; Number 11 in FAO Animal Production and Health Guidelines; Food and Agriculture Organization of the United Nations: Rome, Italy, 2012.
26. ICAR. *Appendix 2 of Section 5 of the ICAR Guidelines: The Standard Trait Definition for Dual Purpose Cattle*; Version March 2022; International Committee for Animal Recording: Rome, Italy, 2022.
27. Yang, G.; Qiao, Y.; Deng, H.; Shi, J.Q.; Song, H. One-stage keypoint detection network for end-to-end cow body measurement. *Eng. Appl. Artif. Intell.* **2025**, *146*, 110333. [CrossRef]
28. Huang, J.; Zhu, Z.; Guo, F.; Huang, G. The Devil Is in the Details: Delving Into Unbiased Data Processing for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5699–5708.
29. Caron, M.; Touvron, H.; Misra, I.; J'egou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 9630–9640.
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
31. Fischler, M.A.; Bolles, R.C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* **1981**, *24*, 381–395. [CrossRef]
32. Fitzgibbon, A.; Pilu, M.; Fisher, R.B. Direct Least Square Fitting of Ellipses. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 476–480. [CrossRef]
33. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision—ECCV 2014*; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.
34. Li, K.; Teng, G. Study on Body Size Measurement Method of Goat and Cattle under Different Background Based on Deep Learning. *Electronics* **2022**, *11*, 993. [CrossRef]
35. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
36. Jacob, B.; Kligys, S.; Chen, B.; Zhu, M.; Tang, M.; Howard, A.G.; Adam, H.; Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2704–2713.
37. Rajafillah, C.; El Moutaouakil, K.; Patriciu, A.M.; Yahyaouy, A.; Riffi, J. INT-FUP: Intuitionistic Fuzzy Pooling. *Mathematics* **2024**, *12*, 1740. [CrossRef]