

A CEDNet-based Algorithm for Cattle Detection and Counting in Complex Farming Environment in UAV Imagery

Tao Wu

Chengdu University of Information Technology

Yihan Niu

Chengdu University of Information Technology

Xi Chen

Southwest Minzu University

Qizhao Zhou

zhouqz@cui.t.edu.cn

Chengdu University of Information Technology

Research Article

Keywords: Cattle Detection, Cattle Counting, Computer Vision, Point Annotation, Smart Livestock

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8272177/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

A CEDNet-based Algorithm for Cattle Detection and Counting in Complex Farming Environment in UAV Imagery

Tao Wu^a, Yihan Niu^a, Xi Chen^c, Qizhao Zhou^{a,b,*}

^a (School of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China)

^b (State Key Laboratory of Hydraulics and Mountain River Engineering, Chengdu, Sichuan 610065, China)

^c (School of Computer Science, Southwest Minzu University, Chengdu, Sichuan 610041, China)

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).

Abstract

UAV imagery has emerged as an efficient modality for large-scale cattle farming, yet existing methods struggle under severe occlusion, cluttered backgrounds, and the high cost of dense annotations. For efficient deployment on edge computing devices with limited processing power, we proposed a new lightweight algorithm for cattle detection named C2Net based on an enhanced CEDNet. By introducing the Focus Inverse Distance Transform (FIDTM) as a supervisory signal, C2Net successfully simplifies the costly requirement of bounding box or polygon annotations to low-cost point annotations, significantly lowering the threshold for data preparation. We also propose our self-constructed dataset HerdSet7K, which is a scale-invariant proximity metric based on Minimum Spanning Trees (MST) together with an object-saliency index to quantify the difficulty of cattle farming scene. The results showed that C2Net high-precision counting (comprehensive MAE: 20.68) while maintaining accurate individual localization capability of those required for bounding-box labels, offering a high-accuracy, low-cost, and practical solution for UAV-based automated livestock monitoring.

Keywords: Cattle Detection; Cattle Counting; Computer Vision; Point Annotation; Smart Livestock

1 Introduction

Livestock farming dominates global land use through direct grazing and indirect feed consumption, contributing approximately 40% of the world's total agricultural output value (FAO, 2015). Within this sector, cattle farming stands as the most significant industry. Countries like the United States (USDA, 2023) and Australia (ABARES, 2023) have long maintained the largest share of cash income from agricultural products through cattle farming. Cattle possess high individual value and require large-scale breeding operations, involving numerous breeding factors that make them representative of the livestock industry. These characteristics have consistently made cattle farming a key focus in smart agriculture research. Automating and unmanned operations in cattle breeding processes to enhance production efficiency and reduce costs remains a crucial research objective in smart agriculture.

This paper focuses on object detection and counting technology in various breeding scenarios of cattle farming, which can provide precise data support for breeding management, facilitate rational planning of breeding scale and resource allocation. As depicted in the top row of Figure 1, the utilization of computer-vision algorithms for processing UAV imagery has been demonstrated to be both feasible and practical for the detection and enumeration of cattle in production environments. (Shao et al., 2019) utilized the YOLOv2 algorithm for real-time cattle detection and counting in drone-captured imagery. Subsequent research has incorporated factors such as camera tilt and image resolution: (Barbedo et al., 2020) employed an Xception-based Convolution Neural Network (CNN) to investigate detection accuracy under oblique viewing angles and varying image resolutions.

Instance-segmentation techniques have also been applied to cattle counting. (Xu et al., 2020) and (Bello and Oladipo, 2024) partitioned their experimental scenes into confined feedlots and open rangeland, testing Mask R-CNN and Mask YOLOv7, respectively. Both studies compared two annotation strategies—head-only versus full-body labels—and found that full-body annotations yielded the highest accuracy in both settings, whereas head-only annotations offered a better trade-off between performance and annotation cost. (de Lima Weber et al., 2023) examined the impact of flight altitude, training and testing YOLOv4 and YOLOv5 variants (sizes X, L, M, S) on images captured at heights from 20 m to 100 m. All configurations achieves above 90 %, with some exceeding 98 %.

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).



Fig.1. Multiple algorithms used in SOTA (abcd), actual unexplored problems (efgh), different behaviors in different scenarios (ijkl).

Although YOLO has a faster detection speed than two-stage detection algorithms, it still retains a large number of parameters and computations in its network because YOLO was proposed for generic multicategory scenarios. (Soares et al., 2024) used the Ford-Fulkerson graph algorithm to detect and eliminate duplicate cattle across multiple attributes, enabling UAV-based computer-vision systems to coherently and efficiently inventory cattle across expansive rangelands. They further emphasize the feasibility and potential of incorporating drones with sophisticated vision algorithms for extensive livestock surveillance. While existing studies have achieved a certain degree of cattle object recognizing accuracy, While prior studies have established a crucial foundation for cattle detection and counting in UAV imagery, several significant limitations remain:

(1) **Inadequate Handling of Spatial Scale:** State of the art(SOTA) approaches fail to address key spatial scale challenges prevalent in real-world cattle farming scenarios adequately, including cattle crowding scenarios, complex background scenarios, significant individual size variations and uneven distribution patterns (second row of Fig. 1.)

(2) **Oversimplification of Farming Scenarios:** The categorization of farming scenarios lacks granularity. As shown in the third row of Fig. 1, the dense cattle aggregation scenarios occur frequently. These result from anthropogenic factors (e.g., fencing and herding; (Negron-Perez et al., 2019)) and spontaneous cattle behavior (e.g., socializing and feeding;(Kondo, 2011)). While existing detection algorithms perform well on sparsely distributed cattle in high-resolution imagery, they exhibit significant performance degradation in such complex scenarios.

(3) **Prohibitive Annotation Costs in Dense Scenarios:** Both SOTA detection models trained on rectangular bounding boxes and segmentation models trained on polygonal masks incur prohibitively high annotation costs within dense cattle herds. Under comparable conditions, the average annotation time for rectangular bounding boxes exceeds that of point-based annotations by more than a factor of four. Furthermore, the annotation cost for polygonal masks can reach up to twenty times that of point annotations. This high cost barrier impedes efficient algorithm training and hinders the broader applicability of these methods to other species.

To address the aforementioned challenges, this paper proposes a novel solution called **C2Net (Cow-CenterNet)** for dense cattle herd counting. Inspired by crowd counting algorithms that predict population density through density maps, C2Net employs FIDTM (Focal Inverse Distance Transform Map) as its supervisory signal. This approach not only maintains counting accuracy but also achieves higher spatial precision, effectively resolving the fundamental issue of density maps being unable to locate

individual cattle. Positioning information serves as the foundation for advanced data analysis, such as monitoring feeding behaviors, herd distribution patterns, and fighting alerts in cattle farming operations. Consequently, C2Net demonstrates broader application potential.

Our main contributions are summarized in the following points:

1) **Dataset Construction.** Images in our cattle farming dataset covering challenging scenarios such as local dense and ultra-dense herds. For the first time, we have integrated and developed the HerdSet7K cattle dataset, providing a high-precision, cost-effective solution for dense cattle detection and counting from drone perspectives. Then, the labeled images were processed for data augmentation to create HerdSet7K. This publicly available HerdSet7K dataset supports the development and benchmarking of advanced cattle behavior analysis.

2) **Novel Network Construction.** We proposed C2Net, which has a smaller scale network than SOTA method. Inspired by crowd counting algorithms that predict population density through density maps, C2Net employs FIDTM as its supervisory signal. This approach not only maintains counting accuracy but also achieves higher spatial precision, effectively resolving the fundamental issue of density maps being unable to locate individual cattle.

3) **New Results.** The results showed that C2Net high-precision counting (comprehensive MAE: 20.68) while maintaining accurate individual localization capability of those required for bounding-box labels, offering a high-accuracy, low-cost, and practical solution for UAV-based automated livestock monitoring. The model demonstrated better performance in terms of both accuracy and speed compared with other lightweight models.

2 Cattle Farming DataSet and Methodology

2.1 Data Acquisition and Processing

UAV-based cattle farming heavily rely on object detection technology, a domain where existing algorithms excel in accuracy but struggle under severe occlusion, cluttered backgrounds, and the high cost of dense annotations. In this paper, a complex cattle farming dataset (HerdSet7k) was constructed, covering challenging scenarios such as local dense and ultra-dense herds. This publicly available dataset supports the development and benchmarking of advanced object detection and cattle behavior analysis. The raw images and videos of the proposed HerdSet7K dataset were sourced from Roboflow(<https://app.roboflow.com/>), the world's largest open-source dataset platform.

To quantify spatial distribution density of targets in variable-resolution images, we developed a scale-invariant proximity metric by calculating the Minimum Spanning Tree (MST) for annotated points in each image. The MST construction follows the classical method (Gower and Ross, 1969), generating a set of $N-1$ edges to connect all points with minimal total length while avoiding circular connections. During implementation, we utilized the optimized graph algorithm from SciPy's spatial module (Virtanen et al., 2020), applying Kruskal's algorithm (Kruskal, 1956) to the complete distance graph to ensure global optimality.

The proximity index is obtained by dividing the median edge length of MST by the diagonal length of the image. This transformation produces a dimensionless measurement that takes into account the variation of resolution and scale differences between different images in essence. The definition of the proximity index is as follows:

$$\text{Proximity Metric} = \text{Median}(\text{MST edge lengths}) / \text{Image Diagonal Length} \quad (1)$$

This computational method produces a scale-invariant metric where lower values indicate denser configurations. Near-zero values signify extremely dense clusters approaching the minimum separation distance, while increasing values reflect progressively sparser distributions. Through empirical analysis of diverse image datasets, three distinct density levels have been established:

- **High density: images with proximity ≤ 0.03**
- **Medium density: images with similarity between 0.03 and 0.05**
- **Images with low density approximation measures > 0.05**

We computed the cattle density of each image using the method described above and collected data in the proportions illustrated in Fig. 2. The collection encompasses two extensive categories, namely confined farming and open grazing, and encompasses intricate sub-scenarios such as beef feedlots, modern large-scale ranches, traditional small-scale ranches, grassland grazing, and woodland grazing.

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).

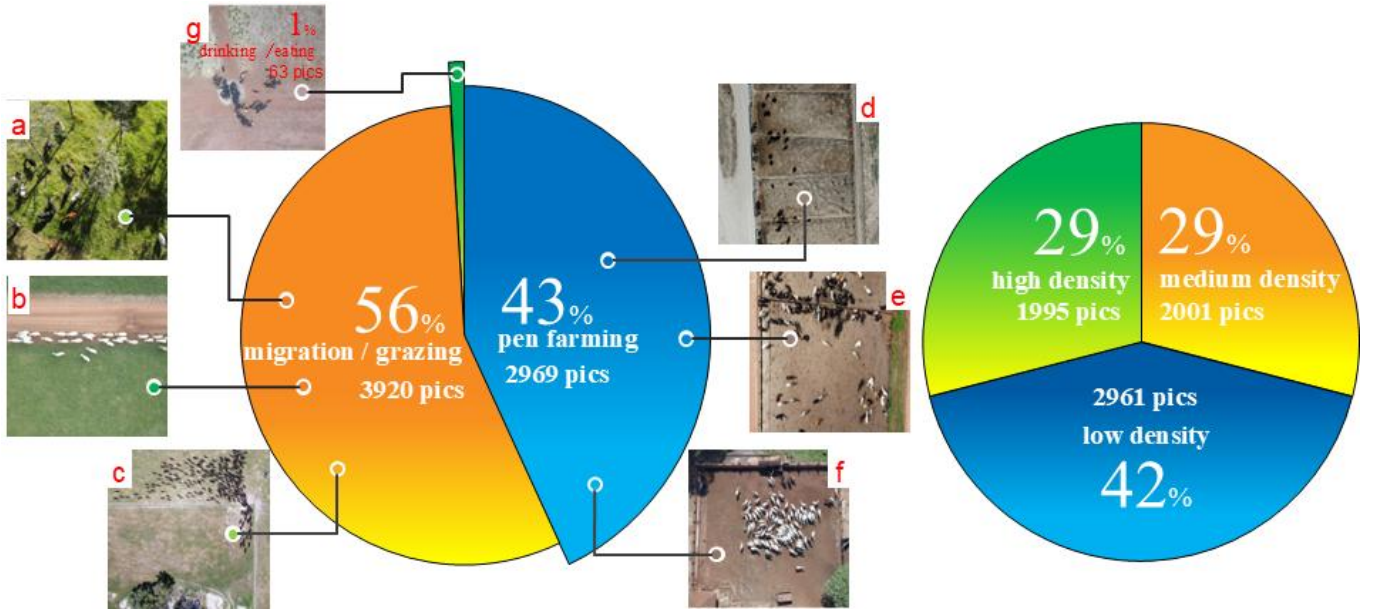


Fig.2. Dataset distribution. As shown in the left pie chart, the entire dataset is divided into three main scene categories: Migration/Grazing, Pen Housing, and Drinking/Feeding. Each category comprises several sub-scenes: (a) woodland grazing, (b) grassland grazing, (c) grassland grazing, (d) pen housing, (e) pen housing, (f) pen housing and (g) feeding. Some pen-housing sub-scenes already contain trough-based drinking/feeding behaviors—e.g., (e)—so the separately listed 1 % Drinking/Feeding slice represents scenes where such behaviors occur completely outside pen housing, as in (g). The density levels of these sub-scenes vary: for instance, both (e) and (f) are pen-housing sub-scenes, yet (e) is moderately dense while (f) is highly dense.

These scenes encompass a wide range of cattle behaviors (drinking, feeding, social interaction, migration) and two camera viewpoints (nadir and oblique). The curation of HerdSet7K raw data is completed by integrating all these factors. HerdSet7K employs more sophisticated partitioning criteria and a more comprehensive scene classification system, thus offering the model a more diverse array of operating environments, enhancing its generalization capacity, and facilitating a comprehensive assessment of its merits and demerits.



Fig.3. The annotation procedures of HerdSet7K dataset

Unlike frontal view crowd counting, cattle in a top-down perspective often do not expose their heads when lowering their heads or being crowded, making head detection insufficient for object recognition. This study employs body center points as representative markers. Figure 3 shows the example of center point offset in HerdSet7k. Larger objects have wider central point ranges. To standardize the positions of large object centers, approximate center points are generated from surrounding bounding boxes before annotation. As shown in Figure 3, bounding box center points may shift due to object pose variations, sometimes even deviating from the object itself. The bounding box annotation method aims to annotate all objects within bounding boxes, with generated center points solely for presence detection. Manual adjustments can align these points to the projection center of 2D images.

In this paper, euclidean distance is employed to calculate the offset distance of each center point, measured in pixels. Each center point requires an offset of approximately 6 pixels on average to achieve uniform relative positioning across all center points. Taking the training set as an example, the overall relative offset between the original center points and the manually adjusted center points is an average offset of 6.7 pixels. This exceeds the test standard's radius threshold of 5 pixels, demonstrating the necessity of manual adjustment. Finally, images and their corresponding annotations are resampled to resolutions compatible with the target architectures: C2Net and DLA-34Seg are trained and tested at 512* 512 px, whereas the YOLO and D-Fine families retain their default 640*640 px resolution.

HerdSet7k consists of 6,952 images and 300,000 instances. According to the division method described in Section 2.1.1, the training set comprises 6,713 images, while comprehensive performance testing is conducted on a test set of 239 images.

2.2 C2Net Algorithm Design

Our proposed C2Net utilizes a CEDNet architecture with an encoder-decoder structure as its backbone network. As shown in Fig. 4, C2Net employs dual-task decoding mechanisms including a localization head and a classification head. The system generates classification maps from the original 512x512 image through a 32x downsampling ratio, producing 16x16 resolution output. After two upscaling processes, it generates a 256x256 FIDT localization map. The network consists of three core components: **the CED backbone network, the classification head, and the localization head.**

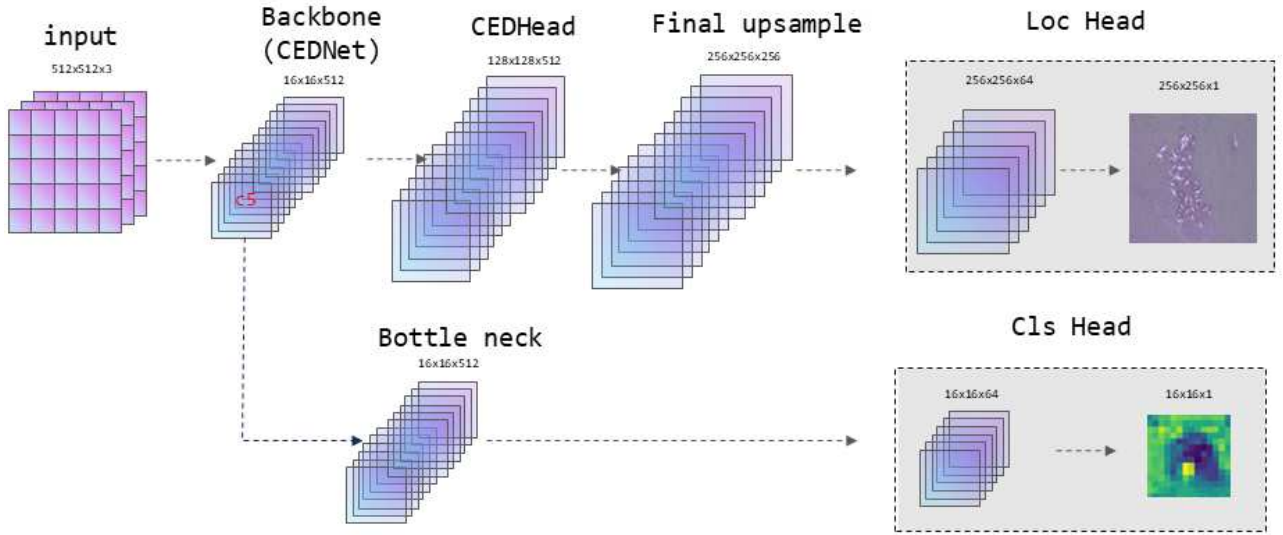


Fig.4. C2Net based on CEDNet architecture with an encoder-decoder structure

2.2.1 CEDNet Backbone Network

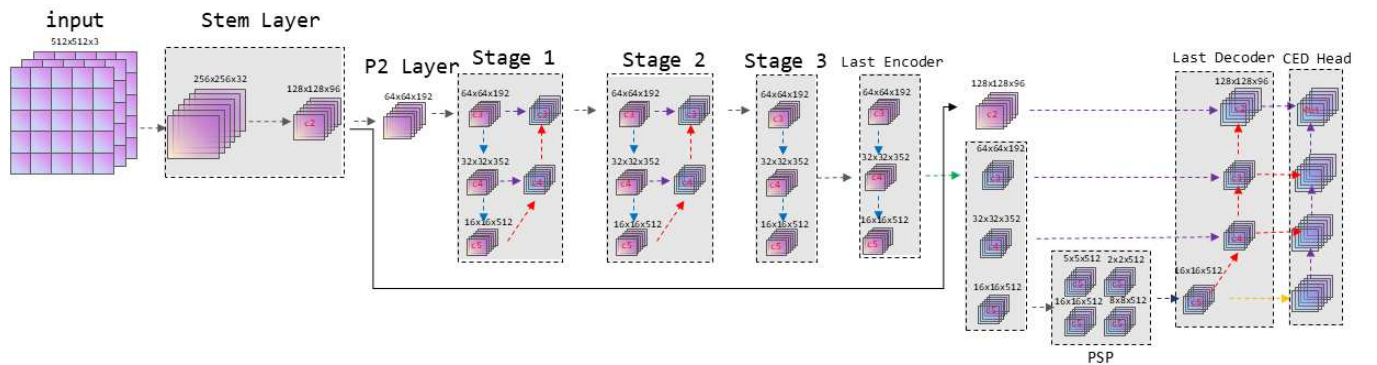


Fig.5. CEDNet architecture

As illustrated in Figure 5, CEDNet is an efficient multi-scale feature extraction backbone network. Inspired by ConvNext and incorporating the successful practices of Vision Transformer (Vit), CEDNet replaces the widely used BatchNorm normalization and ReLU activation functions in classical CNNs with LayerNorm normalization and GeLU activation functions. This eliminates the impact of small batch training data on experiments, making training more stable. CEDNet features two basic convolutional

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).

blocks: the NeXtBlock using 7*7 deep convolution for conventional feature extraction, and the DNeXtBlock employing a dual-path design combining regular convolution and spatially selective convolution to capture broader contextual information.

Distinct from CEDNet's design, the cascade Encoder-Decoder architecture enables progressive feature fusion. Input data first undergoes fourfold downsampling through the Stem Block, then enters a three-stage cascade Encoder-Decoder module. Each stage involves two rounds of downsampling and upsampling for hierarchical feature fusion. This approach allows early-stage integration of high-level features to guide subsequent low-level feature learning, thereby enhancing the effectiveness of multi-scale feature fusion. The deepest feature maps are enhanced with scale context information through a PSP module, and then fused by CED Head to provide multi-scale feature maps for downstream modules.

2.2.2 Positioning Head and Classification Head

The C2Net model employs a decoupled head architecture to achieve joint localization and classification tasks. The localization head utilizes multi-scale fused high-resolution spatial features from the CED backbone network, extracting spatial information through a 64-channel 3*3 convolution layer followed by ReLU activation. It then generates target thermal maps with resolution alignment maintained by a single-channel 1*1 convolution and Sigmoid function, focusing on capturing precise spatial distributions of targets. The classification head (Classification Head) processes low-resolution features from the CED backbone network through a similarly structured 64-channel 3*3 convolution and ReLU activation, directly outputting channel-specific classification feature maps via a 1*1 convolution. This dual-head collaborative architecture innovatively decouples spatial details from semantic features: The localization head enhances small-target sensitivity using shallow high-resolution features, while the classification head improves fine-grained classification capabilities through deep features. Through a spatial alignment mechanism in the feature pyramid, both heads achieve end-to-end joint optimization.

2.3 Focal Inverse Distance Transform Map

Traditional density map regression methods generate density maps as supervised signals through Gaussian kernel convolution. Although they can effectively predict the total number of targets, they have significant defects in high-density areas: overlapping Gaussian responses lead to local maxima that are not aligned with the actual target positions, and they cannot accurately provide individual position and scale information.

To overcome this limitation, C2Net employs the Focal Inverse Distance Transform Map (FIDTM) as a supervisory signal. By applying inverse distance transform, it converts sparse point-level annotations into structured distance representation maps, with the core advantages being:

(1) **Precise Localization Capability:** Generate non-overlapping response regions for each target point. The closer to the target center, the higher the response value (normalized to [0,1]), with adjacent targets having non-overlapping responses. This ensures that local maxima strictly correspond to the target center, effectively resolving the ambiguity issue in Gaussian density maps within high-density regions.

(2) **Scale Perception Modeling:** The response range is controlled by adaptive distance threshold, and the size of the response area is dynamically adjusted according to the local density of the target, which indirectly encodes the individual scale information and provides implicit size supervision for the detection task.

(3) **Point-level Annotation Compatibility:** The FIDTM label can be generated only by the target point-level annotation (the projection coordinates of the center point in the cow top view or side view in this paper), which avoids the dependence on the boundary box annotation and greatly reduces the cost of model training.

First, the formula of Euclidean distance transformation mapping needs to be explained, which is defined as:

$$P(x, y) = \min_{(x', y') \in B} \sqrt{(x - x')^2 + (y - y')^2} \quad (2)$$

Where B denotes the set of all annotations. For any pixel (x, y), (1) means that the pixel value P(x, y) represents the distance between the pixel and its nearest central point (annotation). FIDTM is defined as:

$$I = \frac{1}{P(x, y)^{(\alpha \times P(x, y) + \beta)} + C} \quad (3)$$

Here, I represents the FIDT mapping, with α and β empirically set to 0.02 and 0.75 respectively, and C fixed at 1 to prevent division by zero. FIDT exhibits extremely high response values near the target point, peaking at the center point where the value is set to 1. The response value then rapidly decays as the distance from the center point increases, approaching zero at positions farther away.

2.3 Loss Function

The C2Net positioning branch employs low-resolution FIDTM (FIDT Map) as supervisory signals. C2Net predicts a value for each pixel in the image, collectively forming the predicted FIDTM. The loss function utilizes unnormalized penalty-reduced

pixel-wise logistic regression with focal loss, defined as follows:

$$\mathcal{L}_{\text{loc}}(\hat{Y}_l, Y_l) = - \sum_i \sum_j \begin{cases} (1 - \hat{y}_{l,ij})^\alpha \log(\hat{y}_{l,ij}), & \text{if } y_{l,ij} = 1 \\ ((1 - y_{l,ij})^\beta (\hat{y}_{l,ij})^\alpha \log(1 - \hat{y}_{l,ij}), & \text{otherwise} \end{cases} \quad (4)$$

Here, $\hat{Y}_l$ and Y_l represent the true values and predicted location grids respectively, with $y_{l,ij}$ and $\hat{y}_{l,ij}$ denoting their values at specific pixel positions (i, j) . The focal loss hyper-parameters α and β are set to 2 and 4. For the classification branch of C2Net, low-resolution classification maps are generated from the equivalent low-resolution animal positions $(x'/32, y'/32)$. The commonly used cross-entropy loss function is employed to train this branch:

$$\mathcal{L}_{\text{class}}(\hat{Y}_c, Y_c) = - \sum_i \sum_j \sum_c w_c y_{ijc} \log(\hat{y}_{ijc}) \quad (5)$$

Here, y represents $Y_c \hat{Y}_c y_{ijc} \hat{y}_{ijc} w_c$ the true value of the Duhé coding, and $\hat{y}$ denotes the predicted classification grid. The values of y and $\hat{y}$ at the specific pixel position (i, j) within category c are denoted as y_i and $\hat{y}_i$, respectively, with category weights represented by w . The total training loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{loc}}(\hat{Y}_l, Y_l) + \mathcal{L}_{\text{class}}(\hat{Y}_c, Y_c) \quad (6)$$

2.4 Model Evaluation

In this study, we employed both localization and counting metrics to evaluate trained network architectures. Predictions were classified as true positives (TP) when they matched the ground truth and correctly identified animals. When multiple predictions satisfied both criteria, the best match was selected, while others were categorized as false positives (FP). Finally, cases without matching predictions or incorrect identifications were classified as false negatives (FN). In our C2Net method, we used Euclidean distance between points with a maximum threshold of 5 pixels, where the best prediction was determined by the minimum Euclidean distance. We subsequently calculated recall ("recall"), precision, and F1score for each category (i.e., species) and binary scenario (animal vs. background):

$$\text{recall} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FN}} \quad (7)$$

$$\text{precision} = \frac{\sum \text{TP}}{\sum \text{TP} + \sum \text{FP}} \quad (8)$$

$$\text{F1 score} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \quad (9)$$

Using the mean absolute error (MAE) and root mean square error (RMSE) as counting measures, calculate again for each category and binary situation:

$$\text{MAE} = \frac{1}{I} \sum_{i=1}^I |\hat{n}_i - n_i| \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{1}{I} \sum_{i=1}^I (\hat{n}_i - n_i)^2} \quad (11)$$

Here, I represents the number of images, and $\hat{n}_i, n_i$ respectively represent the predicted number and the actual number of targets in the i -th image.

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).

3 Experimental Results

3.1 Experimental Parameter Settings

The operating system used in our experiment is Linux (Ubuntu 20.04), 12 vCPU Intel (R) Xeon (R). The system features a 2.50 GHz Platinum 8255C CPU and a 32GB vGPU. Python is the programming language, with PyTorch 2.0 as the neural network framework and CUDA 11.8. The training process involves 150 iterations (converging after 77 rounds), 24-batch training, an initial learning rate of 0.001, and Adam as the optimizer.

Table 1 Comparison of detection and counting efficiency

Model	Params(M)	FLOPs(G)
YOLOv8m	25.9	78.9
YOLOv8x	68.2	257.8
YOLO11m	20.1	68.2
YOLO11X	56.97	196.0
DAL-34Seg	18.62	40.52
C2Net(ours)	61.74	157.75

3.2 Comparison Of Detection Performance Of Different Algorithms

We conducted a comprehensive comparison between our method and the most advanced (SOTA) approaches using our dataset, including the YOLOv8 series (YOLOv8m, YOLOv8x) by Sohan et al. (2024), the YOLO11 series (YOLO11m, YOLO11x) by Khanam and Hussain (2024), and DLA-seg as the baseline model for density map-based methods. As shown in Table 2, this evaluation covers key metrics such as detection accuracy (precision P, recall R, F1 score), counting error (average absolute error MAE, root mean square error RMSE), and model efficiency (parameter count Params, FLOPs). In terms of detection accuracy, C2Net demonstrates exceptional comprehensive performance with an F1 score of 0.802, significantly outperforming all comparison models (YOLOv8m:0.744, YOLOv8x:0.780, YOLO11m:0.779, YOLO11x:0.791).

Notably, C2Net achieves the highest recall rate ($R=0.807$), indicating optimal recall while maintaining highly competitive precision ($P=0.798$), successfully achieving the optimal balance between accuracy and recall. Regarding localization accuracy, C2Net shows even more pronounced advantages, with its MAE (20.68) and RMSE (29.94) being the lowest among all models. These values are significantly better than other models (e.g., MAE approximately 36% lower than the second-best YOLO11x and 44% lower than YOLOv8x; RMSE approximately 46% lower than YOLO11x), fully demonstrating its higher precision and stability in target position prediction. Particularly noteworthy is that while achieving the highest accuracy and lowest localization error mentioned above, C2Net maintains moderate levels of parameter size (61.74M) and computational cost (157.75G FLOPs) in terms of model efficiency.

While surpassing lightweight models like YOLO11m(20.1M), YOLOv8m(25.9M) and DAL-34Seg(18.62M) in parameter count, C2Net demonstrates significantly lower computational costs (far below YOLOv8x(257.8G FLOPs), achieving an optimal balance between accuracy and efficiency. Compared to mainstream models, C2Net excels in comprehensive detection accuracy (F1 score) and counting precision (MAE and RMSE), while maintaining optimal recall rates. It also maintains competitiveness in model complexity and computational overhead, delivering a well-balanced performance in accuracy-efficiency trade-offs.

Table 2 Comparison of parameter quantities

Model	P	R	F1score	MAE	RMSE
YOLOv8m	0.885	0.643	0.744	34.61	59.13
YOLOv8x	0.872	0.706	0.780	23.14	39.804
YOLO11m	0.868	0.708	0.779	25.54	42.98
YOLO11X	0.870	0.727	0.791	22.13	33.03
DAL-34Seg	-	-	-	78.38	116.32
C2Net(ours)	0.798	0.807	0.802	20.68	29.94

3.3 Counting Performance Comparison of Different Algorithms

The verification results of all models are shown in Figure 6. C2Net demonstrates the closest match with actual values, fully proving its stable and accurate counting capability across all scales, particularly in large-scale agricultural scenarios where target numbers exceed 500. In such cases, C2Net remains the only model capable of providing usable counting data. Additionally, C2Net's high recall rate enables it to prioritize target presence while maintaining accuracy across various scenarios, a distinctive advantage particularly evident in drone inspection applications where detecting target presence is critical.

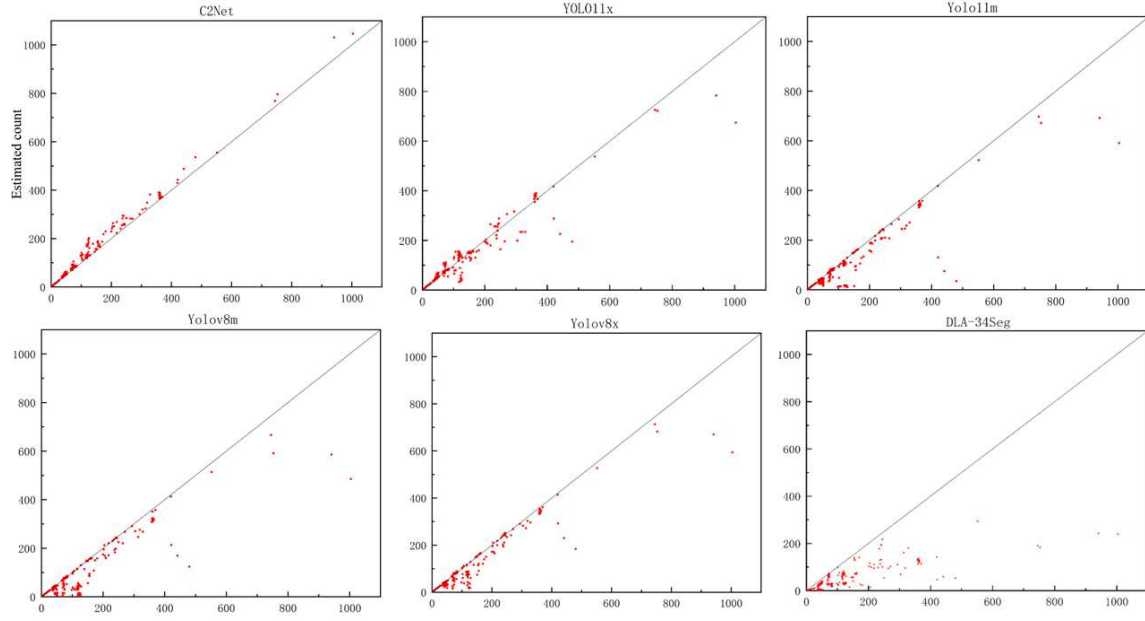


Fig.6. Comparison of counting results for each model on the test set, with the horizontal axis representing the true values, the vertical axis the estimated values, and the trend line $y=x$

3.4 Cow Recognition Effect

To evaluate the performance of the proposed C2Net model in complex and dense cattle herds comprehensively, we conducted detailed comparative experiments with mainstream object detection models. The models included in the comparison were YOLOv8m, YOLOv8x, YOLO11m, and YOLO11x, representing the advanced levels of the YOLO series across different parameter scales. Using our self-built HerdSet7K test dataset, we selected representative scenarios with varying challenges for visual effect comparison and counting accuracy analysis. The specific results are shown in Figures 7, 8, and 9.

As shown in Figure 7, all models face significant challenges in the low-contrast, complex-texture scenarios. These environments present difficulties due to color similarity and texture blending between cattle and backgrounds like muddy surfaces or shadows, which substantially increases the difficulty of distinguishing targets from their surroundings. Experimental results demonstrate that our proposed C2Net model exhibits exceptional robustness in such challenging conditions, with detection and counting performance matching or even surpassing that of larger-parameter models like YOLO11x and YOLOv8x. This indicates that C2Net's strategy combining central point localization with FIDTM supervision demonstrates strong resistance to background interference, enabling more stable detection of target presence.

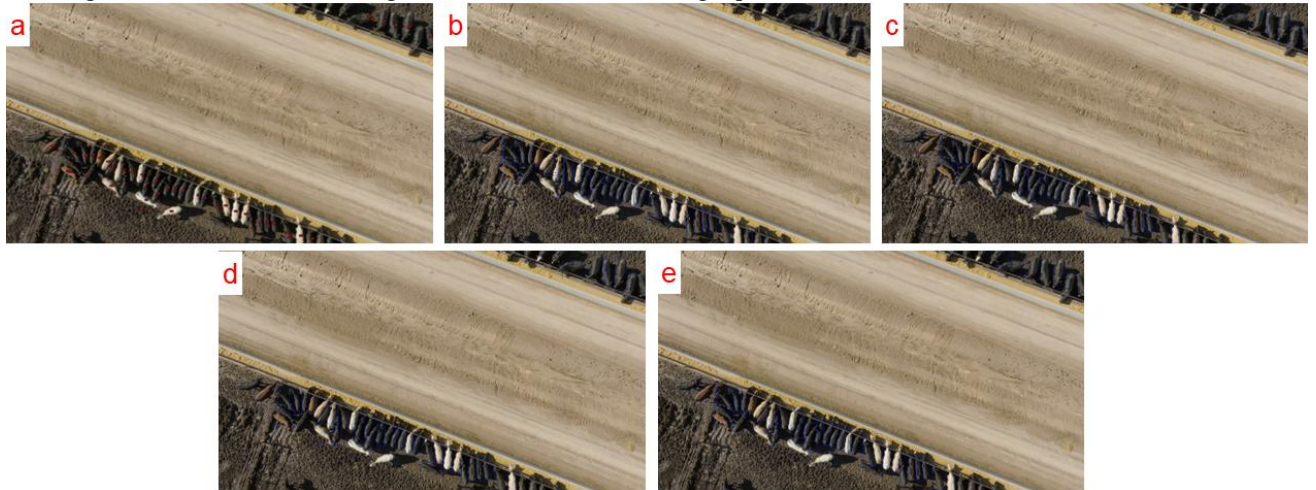


Fig.7. Results of the detection and counting performance of various models in low-contrast complex texture scenes. The figures show: (a) our proposed C2Net, (b) YOLO11m, (c) YOLO11x, (d) YOLOv8m, and (e) YOLOv8x. Our C2Net achieves detection and counting performance that is comparable to or even surpasses that of large-scale models like YOLO11x and YOLOv8x in such scenarios.

*Corresponding author.

E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).

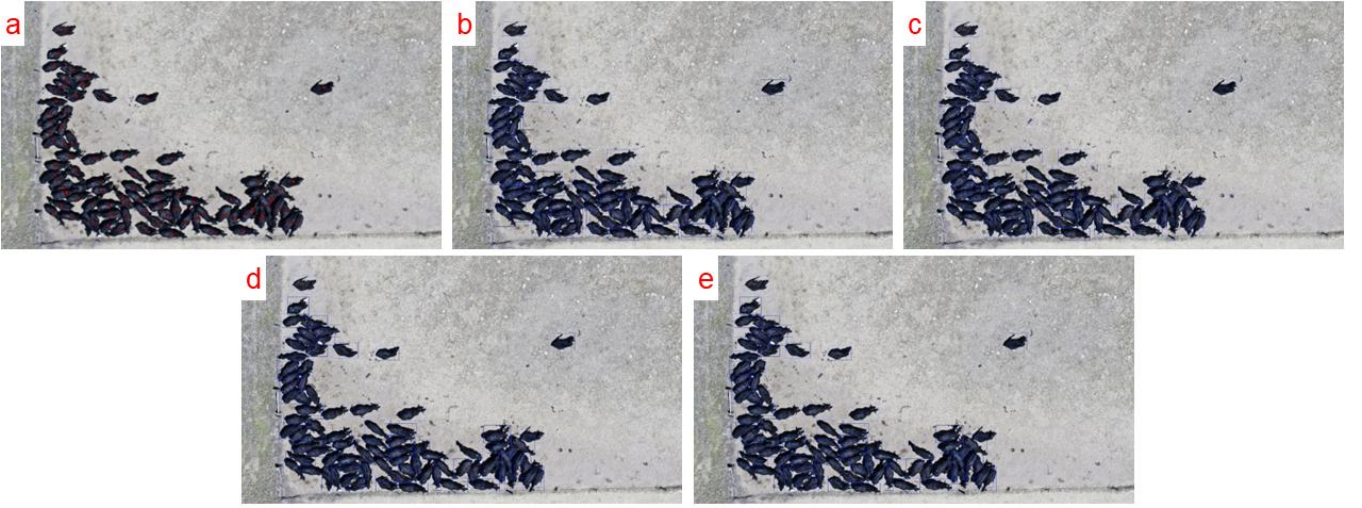


Fig.8. Results of the detection and counting performance of various models in high-contrast simple texture scenes. The models are labeled as: (a) our proposed C2Net, (b) yolo11m, (c) yolo11x, (d) yolov8m, and (e) yolov8x. As shown, all models demonstrate comparable detection and counting performance in this scenario.

In the high-contrast, simple-texture scenarios (as shown in Figure 8), all contrast models demonstrate excellent performance. With high object-background contrast and minimal scene interference, both the YOLO series models based on bounding boxes and our C2Net model achieve high-precision detection and counting, showing comparable performance in these ideal conditions. This confirms that multiple advanced algorithms can effectively solve cattle counting tasks under standard operating conditions.

In the most challenging scenarios featuring low contrast, complex textures, and non-uniform cattle distribution (as shown in Figure 9), the performance of various models exhibited significant divergence. This scenario synthesized the difficulties from the first two scenarios while introducing occlusion and crowding issues caused by uneven spatial distribution. The results clearly demonstrate that YOLO series models suffered from noticeable missed detections and inaccurate counts in such environments, with their detection box integrity also markedly reduced. In contrast, our C2Net model stands out as the only one achieving near-true detection and counting accuracy. This conclusively demonstrates C2Net's unique advantages in handling locally dense and super-dense clusters-common in real-world livestock farming environments. The fundamental reason lies in C2Net's density map regression and precise centroid localization capabilities, which remain insensitive to occlusion and adhesion between targets, enabling better individual resolution within crowded areas.

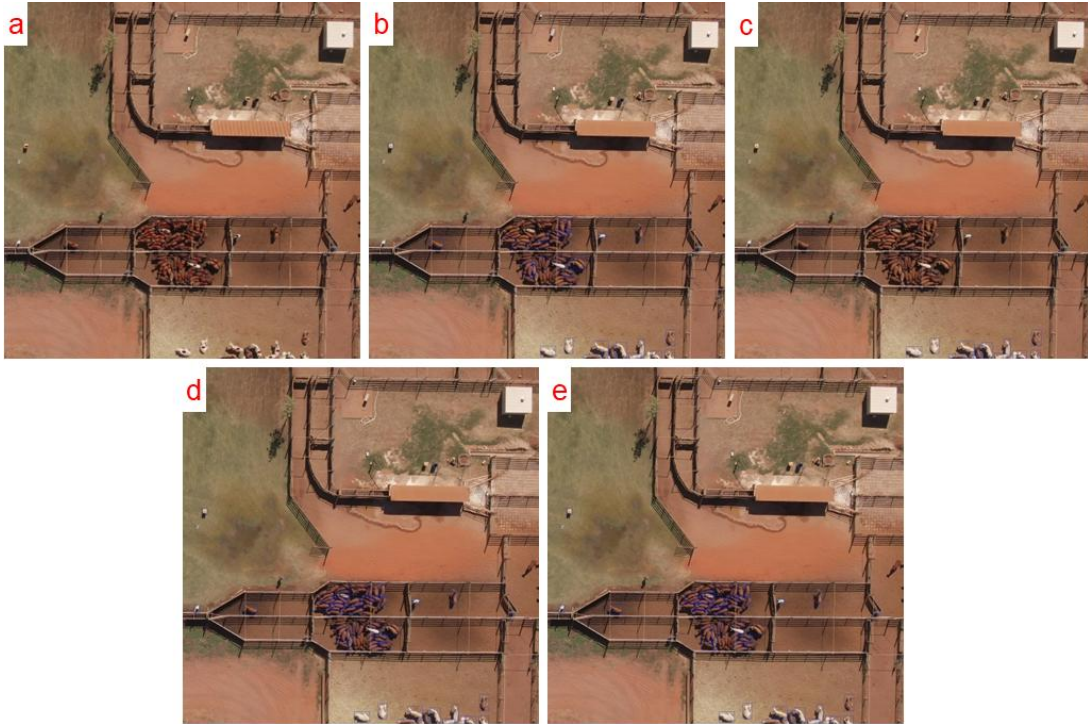


Fig.9. Results of the detection and counting performance of various models under low-contrast complex texture and non-uniform distribution scenarios. The models are labeled as: (a) our proposed C2Net, (b) yolo11m, (c) yolo11x, (d) yolov8m, and (e) yolov8x. As shown, all models exhibit comparable detection and counting performance in this scenario. Notably, our C2Net stands out as the only model achieving detection and counting results that closely

approximate the true values.

Comprehensive analysis reveals that in ideal or moderately challenging scenarios, the proposed C2Net demonstrates performance comparable to mainstream state-of-the-art detection models. However, in real-world farming environments where complex and crowded scenes are more prevalent, C2Net exhibits superior adaptability and robustness, with significantly better counting accuracy and positioning reliability than the comparison models. This addresses the performance degradation issue mentioned in the introduction for existing methods under dense scenarios, validating C2Net's immense potential in practical applications.

4 Conclusions

This study introduces an innovative C2Net object detection framework specifically designed for high-density cattle grazing scenarios. The current challenge is to develop an algorithm that is more accurate than existing cattle detecting methods, which are struggled under complex farming conditions including cluttered backgrounds and the high cost of dense annotations. To address this challenge, a novel paradigm for location-aware counting based on FIDTM supervision signals has been proposed. By introducing the FIDTM as a supervisory signal, C2Net successfully simplifies the costly requirement of bounding box or polygon annotations to low-cost point annotations, significantly lowering the threshold for data preparation. Experimental results demonstrates that, the proposed C2Net achieves the highest recall rate (0.859) and the lowest count error in complex scenarios, meanwhile C2Net achieves high-precision counting (comprehensive MAE: 20.68) while maintaining accurate individual localization capability. In conclusion, C2Net provides a high-precision, efficient, and cost-effective end-to-end solution for dense cattle counting in practical agricultural scenarios. Distinguished by its high-accuracy and low-cost, C2Nets provides a valuable reference for deploying real-time object detection on edge computing devices in cattle framing robots. The model, code, and HerdSet7K dataset are all open-sourced to advance research and development in smart livestock management.

CRedit authorship contribution statement

Tao Wu: Writing - review & editing, Validation, Investigation, Formal analysis, Conceptualization, Funding acquisition. **Yihan Niu:** Writing - original draft, Investigation, Methodology, Data curation. **Xi Chen:** Data curation, Investigation. **Qizhao Zhou:** Writing - review & editing, Supervision, Data curation, Conceptualization, Funding acquisition.

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.

Data availability

Data will be made available on request.

Funding Declaration

This work is supported by the Project of Sichuan Provincial Science Foundation(2025NSFSC2087), and the Open Project of the State Key Laboratory of Hydraulics and Mountain River Engineering(SKHL2413).

References

- FAO,2015. World agriculture:towards 2015/2030 An FAO perspective. Food and Agriculture Organization of the United Nations,URL: <https://www.fao.org/4/y4252e/y4252e07.htm#Ch5>
*Corresponding author.
E-mail address: zhouqz@cuit.edu.cn (Qizhao Zhou).
-

1 USDA,2023. Cattle & Beef - Sector at a Glance. U.S. Department of Agriculture URL:[https://www.ers.usda.gov/topics/animal-products/cattle-](https://www.ers.usda.gov/topics/animal-products/cattle-beef/sector-at-a-glance)
2 [beef/sector-at-a-glance](https://www.ers.usda.gov/topics/animal-products/cattle-beef/sector-at-a-glance)

3 ABARES,2023. Financial performance of livestock farms. Australian Bureau of Agricultural and Resource Economics and Sciences. URL:
4 <https://www.agriculture.gov.au/abares/research-topics/surveys/livestock#daff-page-main>

5 MAHMUD, M. S., ZAHID, A., DAS, A. K., MUZAMMIL, M. & KHAN, M. U. 2021. A systematic literature review on deep learning applications
6 for precision cattle farming. *Computers and Electronics in Agriculture*, 187, 106313.

7 SHAO, W., KAWAKAMI, R., YOSHIHASHI, R., YOU, S., KAWASE, H. & NAEMURA, T. 2019. Cattle detection and counting in UAV images
8 based on convolutional neural networks. *International Journal of Remote Sensing*, 41, 31-52.

9 BARBEDO, J. G. A., KOENIGKAN, L. V. & SANTOS, P. M. 2020. Cattle Detection Using Oblique UAV Images. *Drones*, 4.

10 XU, B., WANG, W., FALZON, G., KWAN, P., GUO, L., CHEN, G., TAIT, A. & SCHNEIDER, D. 2020. Automated cattle counting using Mask R-
11 CNN in quadcopter vision system. *Computers and Electronics in Agriculture*, 171.

12 BELLO, R.-W. & OLADIPO, M. A. 2024. Mask YOLOv7-Based Drone Vision System for Automated Cattle Detection and Counting. *Artificial*
13 *Intelligence and Applications*, 2, 115-125.

14 DE LIMA WEBER, F., DE MORAES WEBER, V. A., DE MORAES, P. H., MATSUBARA, E. T., PAIVA, D. M. B., DE NADAI BONIN GOMES,
15 M., DE OLIVEIRA, L. O. F., DE MEDEIROS, S. R. & CAGNIN, M. I. 2023. Counting cattle in UAV images using convolutional neural network. *Remote*
16 *Sensing Applications: Society and Environment*, 29.

17 SOARES, V. H. A., PONTI, M. A. & CAMPELLO, R. J. G. B. 2024. Multi-attribute, graph-based approach for duplicate cattle removal and counting
18 in large pasture areas from multiple aerial images. *Computers and Electronics in Agriculture*, 220.

19 NEGRON-PEREZ, V. M., FAUSNACHT, D. W. & RHOADS, M. L. 2019. Invited review: Management strategies capable of improving the
20 reproductive performance of heat-stressed dairy cattle. *J Dairy Sci*, 102, 10695-10710.

21 KONDO, S. 2011. Recent progress in the study of behavior and management in grazing cattle. *Anim Sci J*, 82, 26-35.

22 YU, F., WANG, D., SHELHAMER, E. & DARRELL, T. Deep layer aggregation. *Proceedings of the IEEE conference on computer vision and*
23 *pattern recognition*, 2018. 2403-2412.

24 LU, W., CHEN, S.-B., LI, H.-D., SHU, Q.-L., DING, C. H., TANG, J. & LUO, B. 2025. Legnet: Lightweight edge-Gaussian driven network for low-
25 quality remote sensing image object detection. *arXiv preprint arXiv:2503.14012*.

26 LIANG, D., XU, W., ZHU, Y. & ZHOU, Y. 2022. Focal inverse distance transform maps for crowd localization. *IEEE Transactions on Multimedia*,
27 25, 6040-6052.

28 GOWER, J. C. & ROSS, G. J. 1969. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C*
29 *(Applied Statistics)*, 18, 54-64.

30 VIRTANEN, P., GOMMERS, R., OLIPHANT, T. E., HABERLAND, M., REDDY, T., COURNAPEAU, D., BUROVSKI, E., PETERSON, P.,
31 WECKESSER, W. & BRIGHT, J. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17, 261-272.

32 KRUSKAL, J. B. 1956. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical*
33 *society*, 7, 48-50.

34 SOHAN, M., SAI RAM, T. & RAMI REDDY, C. V. A review on yolov8 and its advancements. *International Conference on Data Intelligence and*
35 *Cognitive Informatics*, 2024. Springer, 529-545.

36 KHANAM, R. & HUSSAIN, M. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

37 PENG, Y., LI, H., WU, P., ZHANG, Y., SUN, X. & WU, F. 2024. D-FINE: Redefine regression task in DETRs as fine-grained distribution refinement.
38 *arXiv preprint arXiv:2410.13842*.

39 LI, H., GU, Z., HE, D., WANG, X., HUANG, J., MO, Y., LI, P., HUANG, Z. & WU, F. 2024. A lightweight improved YOLOv5s model and its
40 deployment for detecting pitaya fruits in daytime and nighttime light-supplement environments. *Computers and Electronics in Agriculture*, 220, 108914.

41 CHU, P., LI, Z., ZHANG, K., CHEN, D., LAMMERS, K. & LU, R. 2023. O2RNet: Occluder-occludee relational network for robust apple detection
42 in clustered orchard environments. *Smart Agricultural Technology*, 5: 100284, 2023.

43 PARTEL, V., KAKARLA, S. C. & AMPATZIDIS, Y. 2019. Development and evaluation of a low-cost and smart technology for precision weed
44 management utilizing artificial intelligence. *Computers and electronics in agriculture*, 157, 339-350.

45 DANG, F., CHEN, D., LU, Y. & LI, Z. 2023. YOLOWeeds: A novel benchmark of YOLO object detectors for multi-class weed detection in cotton
46 production systems. *Computers and Electronics in Agriculture*, 205, 107655.
