A Two-Stage Data Efficient Framework for Coat Pattern-Based Cattle Detection and Identification

Meghna Luthra[1], Meghna Sharma[2], Poonam Chaudhary[3]

[1]Department of Computer science & Engineering, The NorthCap University Gurugram, India
[2]Department of Computer science & Engineering, The NorthCap University Gurugram, India
[3]Department of Computer science & Engineering, The NorthCap University Gurugram, India

Abstract
The ability to accurately detect and identify individual cows is essential for precision livestock farming (PLF). It makes monitoring more efficient and provides the foundation for managing diseases, improving productivity, and reducing unnecessary human interference. Conventional techniques of cow detection and individual identification include ear tagging, branding, and visual recognition of cows by humans. However, they can be laborious, intrusive, or inaccurate. In light of the above problem, this research proposes an intelligent cattle detection and identification system with the application of YOLOv8n and transformer architectures, which include ViT (Vision Transformer), DeiT (Data-efficient Image Transformer), and BEiT (Bidirectional Encoder Representation from Image Transformer). Specifically, a two-step intelligent framework is developed for the purpose of automated cattle detection using YOLOv8n and identification using transformer-based algorithms based on visual characteristics and coat patterns. The experimental assessment of the intelligent framework is carried out on the basis of OpenCows2020 dataset, which includes diverse photos taken in various lighting conditions and positions. As a result, excellent cattle recognition is observed, in particular, the most accurate recognition (99.80%) is demonstrated by DeiT, with second place taken by ViT (99.79%), and third by BEiT (95.97%). The detection algorithm demonstrated robust cattle localization (precision 0.993, recall 0.980.

Keywords: Cattle Identification, Cattle Detection, YOLOv8n, Vision Transformers, Precision Livestock Farming, Deep Learning, Automated Livestock Monitoring

1. Introduction

PLF (Precision Livestock Farming) is considered a revolutionizing technique to contemporary livestock management that includes the application of sensing devices, automation, and artificial intelligence technology in order to improve animal welfare, optimize operations and enhance productivity [1]. Cattle identification and accurate monitoring are vital requirements in modern livestock management to improve their welfare, disease management, breeding process, prevent thefts, and traceability [2,3]. Traditional identification methods used for cattle including ear tagging, RFID (Radio Frequency Identification), branding as well as visual inspection. Unfortunately, such conventional methods have several disadvantages such as wear out, lost tags, discomfort for animals, high costs of implementation, dependence on labor, and susceptibility to human errors especially on a large-scale farm [2].

Recent advancements in DL (Deep Learning) and computer vision technology has allowed the creation of automated and non-invasive monitoring systems for farm animals that can overcome these constraints [3]. Deep learning algorithms have achieved impressive results in agricultural automation and biometric recognition of livestock [4]. CNN have been used for computer vision tasks widely being the ability to capture hierarchical features from data and their discriminative power [5]. Various studies have applied facial structure, muzzle pattern, and coat texture recognition using CNNs to identify individual cattle, achieving notable performance levels in agricultural settings [6–8]. Nevertheless, the conventional CNN framework has encountered some challenges in agricultural conditions that involve changes in light sources, partial occlusion, different scales of cattle appearance, complicated backgrounds, and high visual similarity between cattle [3]. Transformers have revolutionized the realm of computer vision through the use of self-attention techniques allowing modeling long-range dependency relationships in an image [9]. While CNNs concentrate on localized features through the receptive field mechanism, Vision Transformers (ViTs) learn features in a global fashion by processing patches of images [10]. Such a characteristic of transformers helps in making transformer-based architectures highly proficient in recognizing objects with high precision, and cattle recognition is one such use case which requires recognizing small differences in the texture and appearance of the animals. Transformer architectures like DeiT [11] and BEiT [12], advanced versions of transformers, have made transformers more efficient in terms of efficiency, representation learning, and classification capabilities.

On the other hand, real-time object detection models have seen much development over time, especially YOLO-based detectors [13,14]. Among these, YOLOv8 stands out to be one of the advanced architectures of YOLO models due to its use of anchor free detection algorithms, better feature extraction methods, efficient localization, and improved performance [15]. Such characteristics make YOLOv8 an ideal choice for implementing livestock monitoring systems. However, the integration of efficient cattle detection algorithms along with accurate and fine-grained cattle recognition

in one intelligent framework still constitutes a research challenge. Most prior work has been dedicated only to cattle detection [4,13,14] or to cattle recognition [6–8]; few works have been devoted to end-to-end intelligent frameworks able to tackle the problems of detection and recognition of cattle simultaneously. Moreover, the solution should be practical by being reliable, scalable, fast, and robust to environmental changes. In this context, this paper presents a framework for automated cattle detection and recognition using YOLOv8n and Vision Transformers. This intelligent framework uses the YOLOv8n algorithm to detect cattle in real-time while applying transformer architectures such as ViT, DeiT, and BEiT to conduct fine-grained cattle recognition from their coat patterns. The capabilities of this system were evaluated using the OpenCows2020 dataset [17]. The contributions of the research are as follows:

- Designing & implementation of an integrated DL framework for automated cattle detection and individual identification.
- Comparative evaluation of transformer-based architectures, including ViT, DeiT, and BEiT, for fine-grained cattle recognition.
- Performance assessment of YOLOv8n for robust real-time cattle detection under challenging agricultural conditions.
- Analysis of the scalability and practical applicability of intelligent livestock monitoring systems for precision agriculture.

This work contributes toward the advancement of automated livestock intelligence by demonstrating the practical feasibility of transformer-assisted computer vision systems for scalable cattle monitoring.

2. Related Work

Automatic cattle identification and detection have received increasing attention in recent times owing to the rising need for intelligent livestock monitoring systems. The conventional techniques used for cattle identification include tagging, branding, RFID, and manual monitoring; these techniques were prone to problems of scalability, wear and tear, high cost, and inefficiency [16]. This initiates to the development of different automated solutions based on computer vision tasks and deep learning advancements for cattle identification and monitoring. Computer vision-based solutions for cattle identification made use of handcrafted features and biometrics for identification purposes. Biometric techniques, like muzzle pattern recognition, have been known to be quite reliable when it comes to identification based on unique muzzle texture. However, the techniques based on computer vision were subject to various constraints pertaining to image acquisition environments and lighting conditions.

Studies have shown that CNN has been successful in discriminating features in cattle based on faces, muzzles, and coats [19, 20]. For instance, Yao et al. proposed CNN model to recognizing cattle faces, thus resulting in high efficiency in identification [21]. In addition, Yang et al. proposed a convolutional framework for recognizing dairy cows based on their faces, proving the importance of deep feature learning in identifying livestock [22]. However, it is important to note that CNN suffers from poor global context and occlusion sensitivity.

The recent transformer-based architectures have enabled remarkable progress in computer vision tasks. The efficacy of the self-attention based mechanism in handling global contexts and long-range dependencies was confirmed by ViTs [9]. The improvement in efficiency of training transformers was achieved in DeiT architecture through the use of knowledge distillation [11] and in BEiT through the use of the masked image modeling technique [12]. This makes transformer-based architectures highly applicable for fine-grained cattle identification based on their coats, which is characterized by large intra-class variance. Object detection models built upon CNN-based architectures have yielded excellent results in the domain of livestock monitoring. The applicability of deep learning-based cattle detection under the condition of aerial surveillance using UAV images was proven in work performed by Barbedo et al. [4]. The popularity of YOLO-based object detectors has become widely recognized, as a result of a balanced trade-off between accuracy and speed of the approach. Petso et al. utilized both YOLOv3 and YOLOv4 for individual herd detection in UAV-based images [14], while Wang et al. developed YOLOv5-based approach for cattle detection with high localization ability [13]. Although the accuracy of detection has been quite satisfactory, some other issues were still relevant.

Recent advancements include the use of hybrid systems utilizing transformers within deep learning algorithms, which have provided impressive outcomes in biometrics of livestock. For instance, Li et al. developed a multi-feature fusion framework at the decision level for cow identification, combining multiple biometric features to ensure reliability [23]. Recognition frameworks incorporating the use of transformer models have proven effective in fine-grained recognition tasks due to enhanced contextual feature modeling [24–26]. However, current studies mainly concentrate either on cattle detection or cattle recognition as separate operations. Few efforts have been made towards developing intelligent frameworks for simultaneous cattle detection and fine-grained individual identification in real farm environments. Moreover, any real-world application necessitates a system that ensures high efficiency and scalability, along with accuracy. These gaps provide the motivation for the current study. Table 1 is showing the comparative findings of existing approaches.

Table 1. Comparative Review of Existing Cattle Identification and Detection Approaches

Study	Methodological Structure	Task	Strengths	Limitations
Kumar et al. [18]	Muzzle Pattern Recognition	Identification	Unique biometric features	Controlled acquisition required
Yao et al. [21]	CNN-based Face Recognition	Identification	Automated feature extraction	Limited global contextual modeling
Yang et al. [22]	CNN-based Facial Recognition	Identification	Strong classification performance	Sensitive to pose variation
Barbedo et al. [4]	Deep Learning + UAV Detection	Detection	Effective aerial cattle localization	Limited fine-grained identification
Petso et al. [14]	YOLOv3 / YOLOv4	Detection	Real-time performance	Reduced robustness in crowded scenes
Wang et al. [13]	Improved YOLOv5	Detection	High localization accuracy	Computational overhead
Li et al. [23]	Multi-feature Fusion	Identification	Robust multi-biometric learning	Higher implementation complexity
Proposed Study	**YOLOv8n + ViT / DeiT / BEiT**	**Detection + Identification**	**Unified intelligent framework, high accuracy, scalability**	**Computational resource dependency**

3. Proposed Methodological View

3.1 Architectural Design and Framework

This study suggests an intelligent deep learning architecture for automatic cattle detection and individual identification by leveraging real-time object detection combined with transformer-based visual recognition methods. The suggested intelligent framework leverages the lightweight nature of YOLOv8n object detection model in conjunction with transformer-based visual recognition methods to develop a comprehensive livestock monitoring framework for precision agriculture applications. The proposed intelligent framework included two sequential steps. In the first, cattle are automatically localized within full-scene images using YOLOv8n object detection architecture, which predicts accurate bounding boxes around detected cattle instances. As a result, this allows isolating cattle regions while eliminating extraneous elements and noise that can negatively impact the process of visual recognition.

In the second step, isolated cattle regions are then classified by transformer-based classification models to detect individual cattle identity using unique characteristics such as coat appearance patterns and structural appearances. Given that the OpenCows2020 dataset consists of 46 unique cattle identities, it means that the recognition task must be devised as a 46-class recognition problem. Specifically, the prime objective is to assign appropriate identity labels to each cattle detected within the image. Thus, the suggested two-step approach allows localizing and isolating cattle first before recognizing their identity.

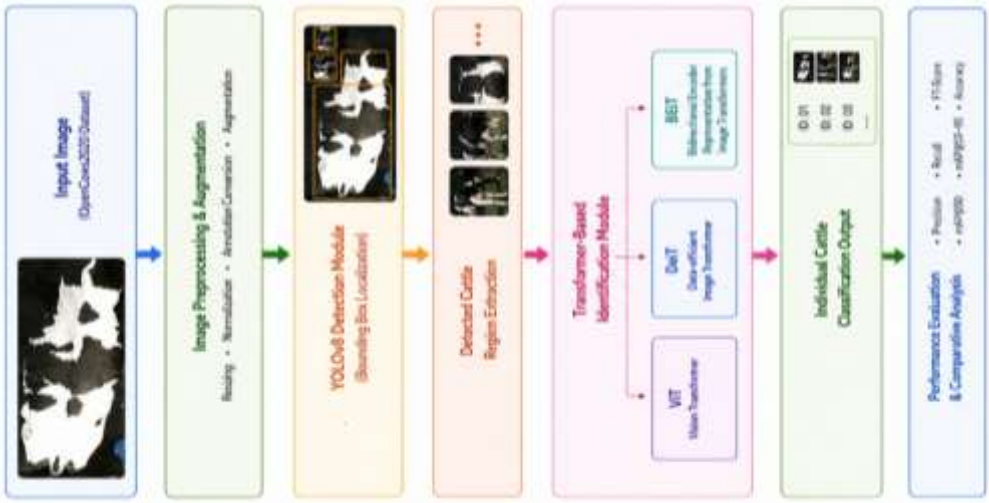

Figure 1. Proposed two-stage intelligent framework for automated cattle individual detection and identification.

3.2 Dataset Description

The experiment was done based on the OpenCows2020 benchmark dataset, which is open-source, used for the detection, localization, and identification of cattle [17]. The summary of this benchmark dataset is mentioned in Table 2.

Table 2. OpenCows2020 Dataset Summary

Parameter	Value
Dataset Name	OpenCows2020
Total Images	11,779
Annotated Objects	13,026
Number of Classes	46
Cattle Breed	Holstein Friesian
Scene Types	Indoor / Outdoor
Environmental Variability	Occlusion, lighting, pose variation

The OpenCows2020 dataset for detection tasks was used to perform cattle localization experiments. Images and Darknet annotation files related to images were imported and synchronized before the experiments started. For data partitioning, the ratio of 80/20 for training/test datasets with random seed equal to 42 was selected through the Equation 1.

$$D = D_{train} \cup D_{test} \quad (1)$$
$$| D_{train} | = 0.8D$$
$$| D_{test} | = 0.2D$$

3.3 Data Preprocessing and Augmentation

The following data pre-processing techniques were used for the robustness and generalizability enhancement of the models: Preprocessing was performed on images including resizing, normalization, label format conversion, and pixel standardization. In case of classification using transformer models, the following input image sizes were standardized:

- **ViT / DeiT input size:** 224 × 224
- **BEiT input size:** 384 × 384

Data augmentation techniques were employed to simulate practical environmental variability encountered in real-world livestock monitoring environments. The augmentation pipeline including Random cropping, Horizontal flipping, Random rotation, Contrast adjustment, Gaussian noise injection, Mosaic augmentation, Color jittering.

The transformations contribute to better robust model to changes in poses, occlusions, lighting, and challenging agricultural environments. As shown in Figure 2, some examples of augmented images that have been created throughout the process of training are presented. The transformations contribute to better visual variety and prevent overfitting by allowing the model to be exposed to more realistic environmental changes.

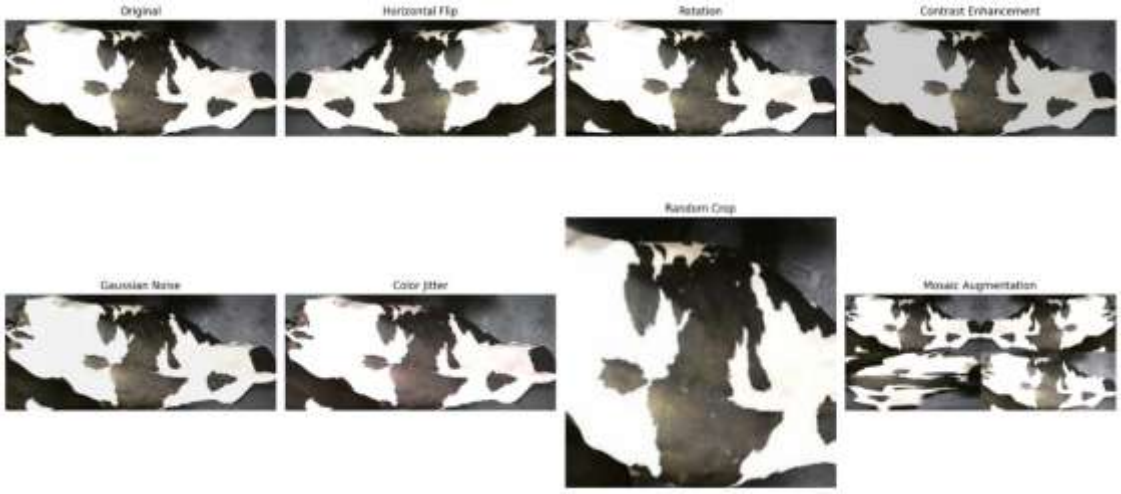

Figure 2. Data Preprocessing & Augmentation Results

3.4 YOLOv8n-Based Cattle Detection Module

The suggested detection system will use the YOLOv8n algorithm for object detection and cattle localization since it is a lightweight neural network, has real-time inference capacity, uses an anchor-less detection scheme, and improves localization in complex agricultural settings [15]. YOLOv8n, as compared to its predecessors, shows increased feature extraction efficiency and decreased computation complexity while improving scaling performance. Figure 3 depicts how object detection takes place based on pre-augmented and normalized images of cows. The training process was

started after uploading the OpenCows2020 dataset pictures and annotations in the Darknet format. As a rule, for training, 80% of data is allocated and the remaining 20% is allocated for testing. Random seed was set in order to make the process reproducible. The YOLOv8n architecture is built of three principal components as Neck, Backbone, and Detection Head.

Backbone conducts hierarchical feature extraction via optimized convolutional operations. It gradually acquires the visual features in low-level and semantic features in high-level that are necessary to precisely localize the cattle. Neck gathers multi-resolution feature representations. Information is transmitted effectively in Neck due to which detection is possible regardless of variation in cattle size, lighting, and view. Detection head uses an anchor-free approach. Anchor free approach simultaneously classifies objects as well as regresses bounding boxes.

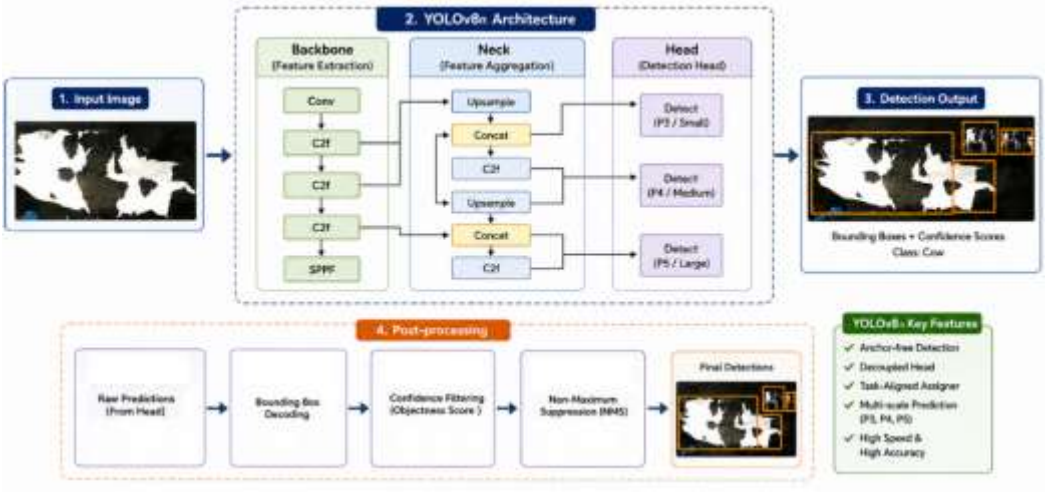

Figure 3. YOLOv8n-Based Cattle Detection Workflow

Training of the model was done with respect to a custom dataset configuration through 300 training epochs. The validation process involved the use of the YOLO evaluation procedure, while the resulting model's weight parameters were saved for future identification tasks. Detection Loss Function is shown in Equation 2:

$$L_{total} = L_{box} + L_{cls} + L_{DFL} \quad (2)$$

where L_{box} is Bounding box regression loss, L_{cls} is Classification loss, L_{DFL} is Distribution Focal Loss, Bounding-box regression is computed using Intersection over Union (IoU)-based optimization described in Equation 3:

$$L_{box} = 1 - IoU(B_{pred}, B_{gt}) \quad (3)$$

Where B_{pred} is Predicted bounding box, B_{gt} is Ground truth bounding box. IoU is defined by Equation 4:

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})} \quad (4)$$

After the predictions, additional post-processing steps such as confidence score filtering and Non-Maximum Suppression (NMS) were applied to remove redundant detection results and keep the best cattle detection results. Table 3 is showing the Summary of the training configuration adopted for the YOLOv8 detection module.

Table 3. Summary of the training configuration

Parameter	Value
Detection Model	YOLOv8n
Training Epochs	300
Train Test Split Ratio	80:20
Number of Classes	1
Annotation Format	Darknet
Validation Method	YOLO Validation

The proposed YOLOv8n detection module enables efficient and accurate cattle localization under challenging agricultural conditions characterized by cluttered backgrounds, overlapping animals, environmental variability, and viewpoint diversity. The localized cattle regions are subsequently forwarded to transformer-based identification models for individual cattle recognition and classification.

3.5 Transformer-Based Cattle Identification Module

After cattle localization using the YOLOv8n detection module, each detected cattle instance is extracted as an individual image region and forwarded to transformer-based architectures for fine-grained cattle identity recognition.

Unlike object detection models, which focus on localizing objects within complex scenes, transformer-based classification models perform discriminative recognition on already isolated cattle instances. The objective of this stage is to accurately assign each detected cattle image to its corresponding individual identity class within the OpenCows2020 dataset, which contains 46 unique cattle identities. Therefore, the identification process is formulated as a 46-class fine-grained classification problem, where subtle differences in coat pattern distribution and structural appearance must be learned for reliable recognition. To evaluate transformer effectiveness for livestock biometric recognition, three advanced transformer architectures including ViT, DeiT, and BEiT were comparatively analyzed. Transformer models represent images as sequences of fixed-size patches. The input image can be represented as Equation 5:

$$X \in \mathbb{R}^{H \times W \times C} \tag{5}$$

Where, H is used for Image height, W is for Image width, C is Number of image channels. The image is partitioned into fixed-size patches and transformed using Equation 6:

$$X_p \in \mathbb{R}^{N \times (P^2 \cdot C)} \tag{6}$$

Where P is used for Patch size, N is Total number of patches. Linear patch embeddings with positional encoding are generated using Equation 7:

$$Z_0 = [x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \tag{7}$$

The self-attention mechanism is defined in Equation 8:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

Where Q shows Query matrix, K shows Key matrix, V is Value matrix, d_k is Key dimension. Classification optimization is performed using Cross-Entropy Loss function shown in Equation 9:

$$L_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{9}$$

Where $\mathbf{y_i}$ represents Ground truth label, $\hat{\mathbf{y}}_\mathbf{i}$ represents Predicted probability. The ability of the Transformer architecture to facilitate global context learning via its self-attention mechanism makes it possible to discriminate accurately between similar looking cattle based on small variations in their coat pattern and physical appearance. This is especially vital in fine-grained animal biometrics identification involving visual similarities between multiple cow identities.

3.6 Experimental Configuration

All experiments were implemented using PyTorch and transformer libraries in GPU-enabled environments. Complete detail of experimental configuration is represented in Table 4.

Table 4. Experimental Hyperparameter Configuration

Parameter	Value
Framework	PyTorch
Detection Model	YOLOv8n
Classification Models	ViT, DeiT, BEiT
Optimizer	AdamW
Weight Decay	0.05
Learning Rate	2×10^{-4}
Batch Size	64
Classification Epochs	5
Detection Epochs	300
Hardware	NVIDIA A100

4. Experimental Results and Performance Discussion

The results of proposed intelligent system for cattle detection with YOLOv8n and cattle identification with transformers (ViT, DeiT, and BEiT) are presented in this section. The system is tested on the basis of standard metrics related to object detection and object classification in order to determine its performance for localization and identification under real-world agricultural conditions. The experiment includes two main parts, namely (i) cattle detection with YOLOv8n and (ii) cattle identification with transformer models.

4.1 Evaluation Metrics

For the evaluation of the YOLOv8n cattle detector model, recall, precision, F1-Score, and mAP (mean average precision) have been employed. Precision is the ratio of true cattle detection cases among all predicted cases are calculated using Equation 10.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Where TP (True Positive) detections, FP (False Positive) detections. High precision indicates fewer incorrect cattle detections. Recall evaluates the ability of the model to detect actual cattle instances (Shown in Equation 11)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

Where FN (False Negative) detections, Higher recall reflects improved detection completeness. F1-Score (Shown in Equation 12) combines precision and recall into a single balanced measure.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

The mAP (mean Average Precision) measures the detection quality across multiple IoU thresholds using Equation 13:

$$\text{mAP} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \text{A P}_\text{i} \tag{13}$$

Where AP (Average Precision) for each class, N (Number of classes). For the transformer-based cattle identification module, classification performance was evaluated using Accuracy and Cross-Entropy Loss. Accuracy measures the proportion of correctly classified cattle images (Shown in Equation 14):

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{14}$$

Where TN (True Negative) classifications. To optimize classification learning, Cross-Entropy Loss was employed using Equation 15:

$$\text{L} = -\sum y \log(\hat{y}) \tag{15}$$

Where y (Ground truth label), $\hat{y}$ (Predicted probability). These evaluation metrics collectively provide a comprehensive assessment of the proposed framework's capability to perform accurate cattle localization and reliable individual cattle identification.

4.2 YOLOv8n Detection Performance

The performance of YOLOv8n detection algorithm was evaluated in order to determine its capability of accurately detecting cows in different environments. The YOLOv8n was trained using OpenCows2020 dataset and its detection performance tested on various images that included different lighting, viewpoint, scale, background, and partial occlusions. To evaluate the detection algorithm, some of the popular detection evaluation metrics like recall, precision, F1-score, mean average precision @50, and mean average precision @50-95 were used. The results are shown in Table 5.

Table 5. YOLOv8n Detection Performance Results

Metric	Value
Precision	0.993
Recall	0.980
F1-Score	0.986
mAP@50	0.995
mAP@50–95	0.940

Figure 4 shows the behavior of convergence between training and validation for YOLOv8n in cattle detection. The gradual decrease in the loss values during training, such as box regression, classification, and distribution focal losses, demonstrates that the model optimization process is stable. This is also supported by the steady increase in precision, recall, and mAP values at 50 and 50-95.

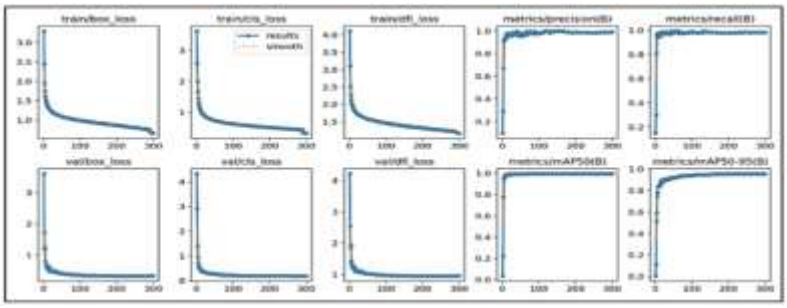

Figure 4. Training and Validation Convergence Behavior of YOLOv8n for Cattle Detection

This similarity between training and validation performance indicates good generalization ability with no visible overfitting, showcasing the high level of robustness of the YOLOv8n architecture. The quantitative analysis shows the highly reliable cattle detection capabilities of the lightweight YOLOv8n. It achieved the highest possible level of precision, reaching 0.993, which means that almost all detections belonged to cattle with no false detections at all. The recall rate equals 0.980, and this metric also shows the great capability of the neural network to detect most of the cattle instances in the data. F1-score equal to 0.986 illustrates the great stability of YOLOv8n with its perfect balance between precision and recall in cattle detection tasks. Moreover, YOLOv8n showed extremely high mAP@50 equal to 0.995, meaning highly accurate localization of cattle with the IoU equal to 50%. Finally, the mAP@50–95 equals 0.940, demonstrating great robustness even with more strict localization criteria. These results prove that the contributions of the anchor-free design, effective feature extraction, and improved localization capabilities of YOLOv8n towards the detection of cattle in challenging agricultural conditions cannot be underestimated. Being able to precisely segment cattle from the scene by identifying individual objects, the developed detection module forms a trustworthy foundation for the further development of the proposed system for cattle tracking.

4.3 Transformer-Based Identification Performance

After the accurate localization of the cattle instances via the YOLOv8n detection model, the selected images were fed into transformer-based networks to recognize individual cattle identity. In this phase, the aim is to match each detected cow with its own corresponding cattle class out of the total 46 cattle classes identified in the OpenCows2020 dataset. Given that cows of the same breed tend to have very similar appearances and coat patterns, recognizing individual cows from each other requires fine-grained biometric classification. Therefore, three transformer architectures named Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), and Bidirectional Encoder Representation from Image Transformers (BEiT) were examined and compared against each other with respect to their ability to recognize individual cattle. The performance of the networks was assessed by using common classification measures like Accuracy, Precision, Recall, and F1-Score shown in Table 6. To be able to make a meaningful comparison, the popular ResNet-50 was used as a convolutional neural network baseline.

Table 6. Comparative Analysis of Deep Learning Models for Cattle Identification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
ResNet-50	98.0%	0.984	0.982	0.982
ViT	99.79	99.81	99.79	99.79
DeiT	99.80	99.80	99.80	99.80
BEiT	95.97	95.77	95.97	95.28

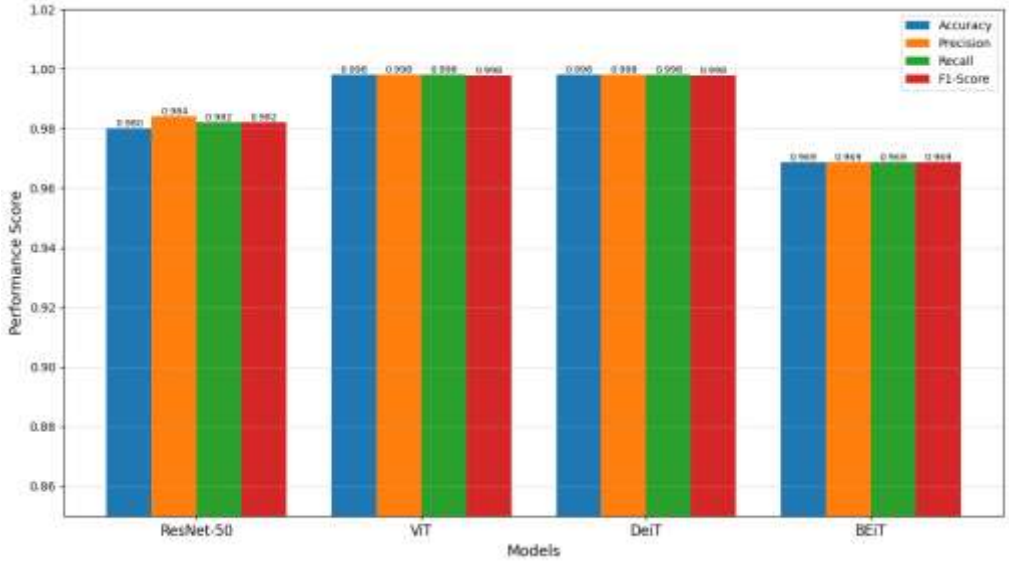

Figure 5. Comparative Analysis of Transformer-Based Cattle Identification Models

Figure 5 highlights the comparative performance analysis of the selected models to validate their performance in cattle identification. All the models were efficient enough in classification and proved to be suitable for cattle biometrics with respect to fine-grained identity recognition using transformers. DeiT provided the highest accuracy of 99.80% compared to other models such as ViT and BEiT. This is mainly due to the data efficiency of DeiT in terms of training and knowledge distillation techniques. Similarly, ViT also showed perfect performance by scoring an accuracy rate of 99.79%. In particular, the performance of ViT in terms of self-attention mechanisms for capturing long-range

dependence proves the effectiveness of the model for fine-grained identification. Similarly, BEiT proved to be successful with an accuracy of 95.97% because of its reliable representation learning technique. Though BEiT performed poorly compared to other models like ViT and DeiT, its efficiency for fine-grained recognition could not be overlooked.

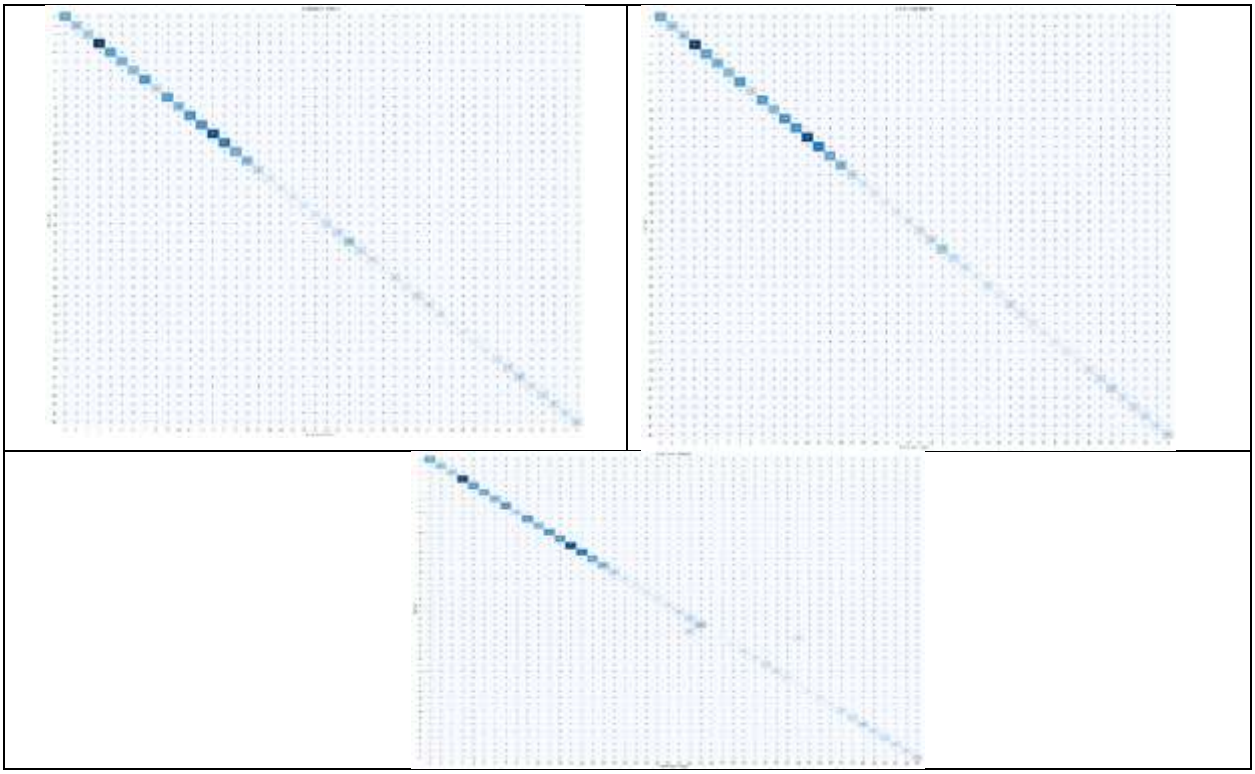

Figure 6. Confusion matrices of transformer-based cattle identification models: (a) Vision Transformer (ViT), (b) Data-efficient Image Transformer (DeiT), and (c) Bidirectional Encoder Representation from Image Transformers (BEiT).

As evident from the confusion matrices depicted in Figure 6, there is an intricate analysis of the classification accuracy for each class using the proposed transformer architectures. There is a depiction of class-wise classification accuracy across 46 cattle identity classes. Both the ViT and DeiT models are characterized by high diagonal dominance with negligible misclassification between different classes, while in case of BEiT, the misclassification rate is higher among visually similar cattle classes. High dominance on the diagonal is also apparent for all the models used, implying that accurate classification is achieved for most cattle identity classes. Diagonal dominance in DeiT and ViT models is almost perfect and with little misclassification, thus confirming their discriminating capabilities for cattle classification. On the other hand, there is significant off-diagonal dispersion in the BEiT model, thus confirming its relatively poor classification capabilities due to high misclassification among visually similar cattle classes.

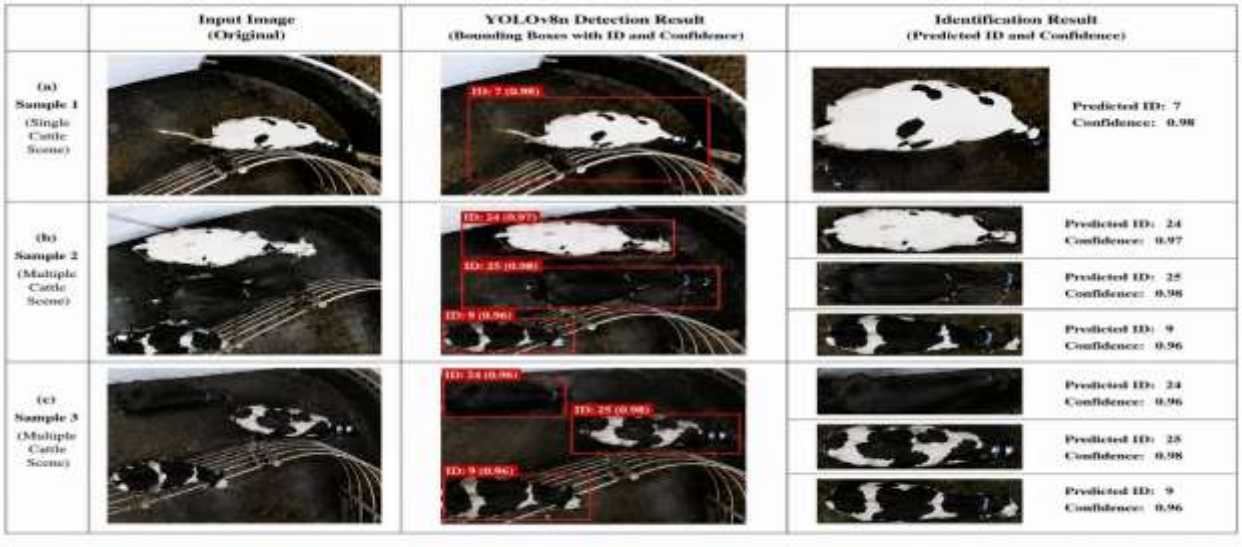

Figure 7. Qualitative End-to-End Cattle Detection and Identification Results

Qualitative results in Figure 7 clearly indicate the successful end-to-end performance of the proposed intelligent cattle monitoring system under practical farming environment. The system effectively performs cattle detection and individual identification in single-cattle and multi-cattle monitoring scenarios. The YOLOv8n detection model effectively localizes the cattle instances, despite challenges in terms of overlapping instances, crowded scenes, and background complexity, which indicates its applicability in real-life livestock monitoring applications. In turn, the localized cattle images are analyzed using transformer-based identification models, which provide the relevant individual identities of the cattle based on the 46 classes defined in the OpenCows2020 database. Indeed, the qualitative results coincide with quantitative performance of the system based on analysis of detection, classification, and confusion matrices in terms of challenging conditions including crowd, different viewpoints, and partial occlusions. Thus, combining YOLOv8n cattle localization and transformer-based individual identity recognition is an efficient approach to automated precision cattle monitoring.

4.4 Comparative Analysis with Existing Approaches

As a part of the validation of the suggested intelligent cattle monitoring system, the comparison with other existing cattle detection and identification methods was carried out using criteria like methodology used, capabilities of each approach, and performance efficiency of the approach in relation to livestock monitoring. The major difference between the suggested intelligent cattle monitoring approach and conventional ones is the fact that the conventional techniques either deal with cattle detection or cattle individual identity recognition whereas the suggested technique combines these two functions within one deep learning approach.

Table 7. Comparative Analysis with Existing Cattle Detection and Identification Approaches

Study	Detection	Identification	Methodology	Overall Performance (%)
Yao et al. [21]	No	Yes	CNN-based cattle face recognition	95.60
Wang et al. [13]	Yes	No	Improved YOLO-based cattle detection	97.40
Li et al. [23]	No	Yes	Multi-feature fusion identification	98.10
Proposed Framework	Yes	Yes	YOLOv8n + ViT / DeiT / BEiT	99.80

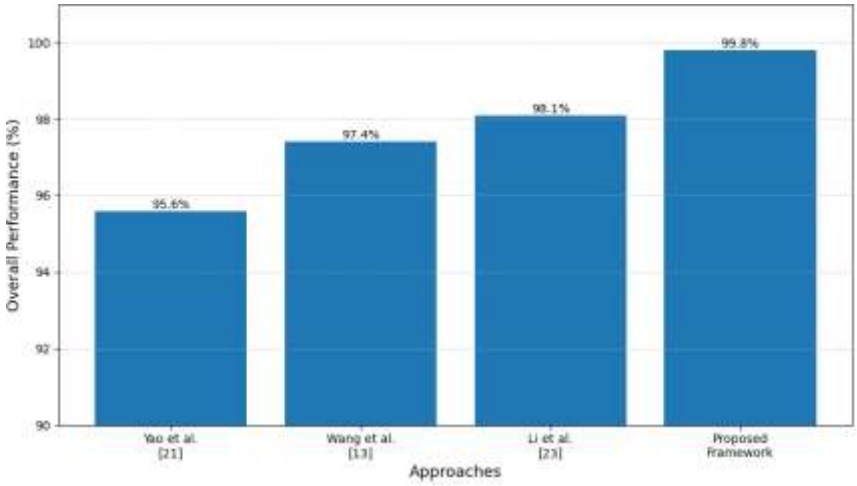

Figure 8. Comparative Performance Analysis with Existing Approaches

As shown in Figure 8 and Table 7, the proposed intelligent cattle monitoring framework outperforms most existing frameworks with regard to cattle detection and individual identity recognition. Existing literature tends to concentrate on one isolated aspect of cattle detection or identification without considering the other. For example, Yao et al. [21] developed an efficient CNN-based cattle recognition framework but could not detect cattle automatically. In addition, Wang et al. [13] designed a high-performance cattle detector based on the improved YOLO model but did not consider cattle individual identification, whereas Li et al. [23] developed a multi-feature fusion cattle recognition algorithm with greater computational requirements. Unlike other methods, our approach combines real-time YOLOv8n-based cattle detection with fine-grained transformer-based individual identity recognition, leading to a highly scalable and practically useful cattle monitoring technique. The above discussion proves the success of our approach and its viability for precise livestock farming purposes. With respect to practical applications, there are various benefits associated with the use of this framework in the field. First, cattle monitoring will be automated. Second, it will be easier for farmers to monitor cattle in a more efficient manner than before, since they will not have to depend on their observations or tags attached.

4.5 Discussion and Practical Implications

The experimental results indicate the efficiency of the proposed deep learning solution for the development of a unified intelligent pipeline for automated cattle identification in precision livestock farming. In combination with YOLOv8n, the application of transformer-based networks made it possible to develop an efficient real-time pipeline combining cattle detection and individual recognition. The localization algorithm based on YOLOv8n showed high localization efficiency in the presence of environmental factors such as different illuminations, complex backgrounds, scale changes, and partial occlusions. High accuracy values are evidence that this algorithm can be used effectively for real-time livestock monitoring purposes. As for identification algorithms, the use of transformer networks allowed classifying objects more efficiently through capturing global context interactions and subtle visual differences between individual objects. Out of all the models, the most efficient was DeiT, followed by ViT. BEiT model was also efficient in terms of cattle recognition. The sequential application of two approaches made it possible to eliminate the problem of background interferences during classification. From an application perspective, the proposed framework offers several advantages for modern livestock management:

- Development of a unified two-stage intelligent framework for automated cattle detection and individual cattle identity recognition.
- Integration of lightweight YOLOv8n for real-time cattle localization under realistic agricultural conditions.
- Comparative evaluation of transformer-based architectures (ViT, DeiT, and BEiT) for fine-grained cattle biometric recognition across 46 unique cattle identities.
- Validation of the framework's scalability and practical applicability for precision livestock monitoring.

Although the framework achieved strong performance, practical limitations remain. Transformer-based models require substantial computational resources, which may restrict deployment in resource-constrained farm environments. Performance may also be affected under severe occlusion, low-resolution imagery, or highly crowded scenes. Future research may focus on lightweight transformer architectures, edge-AI deployment, multimodal sensing integration, and real-time embedded livestock monitoring systems to further improve scalability and operational feasibility.

5. Conclusion and Future Scope

This research introduced a smart and effective deep learning architecture for automated cattle detection and identity recognition using YOLOv8n in combination with different types of transformer-based architectures, namely ViT, DeiT, and BEiT, to resolve the shortcomings of existing livestock identification techniques through an easily scalable, non-invasive, and automated monitoring approach. The introduced architecture used YOLOv8n to detect and localize cattle followed by transformer-based modules to recognize their identities among 46 different cattle classes found in the OpenCows2020 dataset. This strategy minimized the effect of background interference and enhanced cattle identification accuracy. Results revealed highly impressive performance from the model in both detection and identity recognition tasks. For detection, YOLOv8n yielded 0.993 precision, 0.980 recall, 0.986 F1-score, and 0.940 mAP@50-95, proving the ability of the architecture to localize cattle accurately and effectively. In terms of identity recognition, the most accurate model was the DeiT with 99.80%, ViT with 99.79%, and BEiT with 95.97% accuracy, indicating the efficiency of transformer-based self-attention modules in recognizing livestock biometrics in a fine-grained manner. The proposed model combining lightweight object detection with transformer-based identity recognition makes it more feasible to deploy automated monitoring of livestock in agricultural applications.

References

1 D. Berckmans, "General introduction to precision livestock farming," Animal Frontiers, vol. 7, no. 1, pp. 6–11, 2017. DOI:10.2527/af.2017.0102

2 Neethirajan, S. Artificial Intelligence and Sensor Innovations: Enhancing Livestock Welfare with a Human-Centric Approach. Hum-Cent Intell Syst 4, 77–92 (2024). https://doi.org/10.1007/s44230-023-00050-2

3 Guilherme L Menezes, Gustavo Mazon, Rafael E P Ferreira, Victor E Cabrera, Joao R R Dorea, Artificial intelligence for livestock: a narrative review of the applications of computer vision systems and large language models for animal farming, Animal Frontiers, Volume 14, Issue 6, December 2024, Pages 42–53, https://doi.org/10.1093/af/vfae048

4 M. Joshi, C. Pal, S. K. Singh and S. Shrivastava, "Multiscale Reparameterized Deep Architecture for Efficient and Interpretable Livestock Verification," in IEEE Transactions on AgriFood Electronics, doi: 10.1109/TAFE.2026.3671987.

5 Zhi Weng, Fansheng Meng, Shaoqing Liu, Yong Zhang, Zhiqiang Zheng, Caili Gong, Cattle face recognition based on a Two-Branch convolutional neural network, Computers and Electronics in Agriculture, Volume 196, 2022, 106871, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2022.106871.

6 K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

7 Yaowu Wang, Sander Mücher, Kaiwen Wang, Wensheng Wang, Lammert Kooistra, Deep metric learning for individual cattle identification using coat patterns: Proposal for a best practice, Computers and Electronics in Agriculture, Volume 238, 2025, 110754, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2025.110754.

8 Meng, H., Zhang, L., Yang, F., Hai, L., Wei, Y., Zhu, L., & Zhang, J. (2025). Livestock Biometrics Identification Using Computer Vision Approaches: A Review. Agriculture, 15(1), 102. https://doi.org/10.3390/agriculture15010102.

9 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv, abs/2010.11929.

10 B. Palanisamy et al., "Transformers for Vision: A Survey on Innovative Methods for Computer Vision," in IEEE Access, vol. 13, pp. 95496-95523, 2025, doi: 10.1109/ACCESS.2025.3571735.

11 Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & J'egou, H. (2020). Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning.

12 Bao, H., Dong, L., & Wei, F. (2021). BEiT: BERT Pre-Training of Image Transformers. ArXiv, abs/2106.08254.

13 S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017, doi: 10.1109/TPAMI.2016.2577031.

14 J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

15 Bochkovskiy, A., Wang, C., & Liao, H.M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. ArXiv, abs/2004.10934.

16 G. Jocher, A. Chaurasia, and J. Qiu, “Ultralytics YOLOv8,” GitHub repository, 2023. Available: https://github.com/ultralytics/ultralytics.

17 W. Andrew, T. Burghardt, N. W. Campbell, and A. Dowsey, OpenCows2020 Dataset, University of Bristol, 2020.

18 W. Andrew, J. Gao, S. M. Mullan, N. W. Campbell, A. Dowsey, and T. Burghardt, “Visual identification of individual Holstein-Friesian cattle via deep metric learning,” Computers and Electronics in Agriculture, vol. 185, 2021.

19 L. R. Heráldez, L. Barbieri, M. Babuglia and A. Arnaud, "RFID in the Livestock Industry, from Traceability to a Decision Taking Tool in the Cattle-Yards," 2024 IEEE 15th Latin America Symposium on Circuits and Systems (LASCAS), Punta del Este, Uruguay, 2024, pp. 1-4, doi: 10.1109/LASCAS60203.2024.10506117.

20 Liu, S. et al. (2025). Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds) Computer Vision – ECCV 2024. ECCV 2024. Lecture Notes in Computer Science, vol 15105. Springer, Cham. https://doi.org/10.1007/978-3-031-72970-6_3

21 Xiaopu Feng, Jiaying Zhang, Yongsheng Qi, Liqiang Liu, Yongting Li, CATR-Net: Cattle–Attentive transformer with adaptive and enhanced segmentation and recognition, Computers and Electronics in Agriculture, Volume 239, Part B, 2025, 111038, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2025.111038.

22 Swaroop, D., Guru, D.S. (2026). Deep Neural Networks in Cow Face Recognition. In: Zaroliagis, C., Bhandari, D., Gupta, P., Das, S. (eds) Applied Algorithms. ICAA 2026. Lecture Notes in Computer Science, vol 16423. Springer, Cham. https://doi.org/10.1007/978-3-032-15621-1_13

23 Li, D., Li, B., Li, Q. et al. Cattle identification based on multiple feature decision layer fusion. Sci Rep **14**, 26631 (2024). https://doi.org/10.1038/s41598-024-76718-x

24 Liu, Chengyun & Zhao, Feiyang & Huang, Boya & Zhang, Xintong & Zhang, Dequan & Li, Hualin. (2024). Cow face identification based on CNN by using channel attention module and spatial attention module. 25. 10.1117/12.3026338.

25 D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.

26 F. Chollet, Deep Learning with Python. Manning Publications, 2018.