**Journal of Information Systems Engineering and Management**
2025, 10(42s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Cattle Identification and Detection using Vision Transformers and YOLOv8

Meghna Luthra[1], Meghna Sharma[2], Poonam Chaudhary[3]
*[1]Department of Computer Science & Engineering, TheNorthCap University Gurugram, India*
*[2]Department of Computer Science & Engineering, TheNorthCap University Gurugram, India*
*[3]Department of Computer Science & Engineering, TheNorthCap University Gurugram, India*

**ARTICLE INFO**

Received: 24 Dec 2024

Revised: 12 Feb 2025

Accepted: 26 Feb 2025

**ABSTRACT**

Modern livestock management need accurate cattle identification, which enable better monitoring, health tracking, and productivity optimization. The cattle detection locate and detect the presence of cattle within an image or video frame. For cattle detection, this research focuses on YOLOv8 model and for the cattle identification the focus is on use of vision transformers (ViT, DeiT, BEiT). We evaluated the performance of the proposed model using the Opencows2020 dataset. Experimental results indicates that ViT outperformed other models in identification tasks and achieve an accuracy of 99.79%. YOLOv8 effectively detected cattle based on coat patterns that shows its suitability for real-time applications. The findings shows the potential of deep learning in improved cattle management systems.

**Keywords:** Cattle Identification, Vision Transformers, YOLOv8, Machine Learning, Livestock Management

## INTRODUCTION

Livestock management is a key component that facilitates the monitoring, monitoring of health, and productivity maximization. The proper identification and detection of cattle are among the primary challenges under this sector. Identification was carried out conventionally using physical tags or by observing manually. However, conventional practices suffer due to inefficiencies through human errors and practical limitations. To solve these problems, our proposed research examines and experimented the ability of state-of-the-art deep learning techniques, with an emphasis on vision transformers and advanced object detection models. Vision Transformers (ViTs) [1], like DeiT [2] and BEiT [3], are applied to recognize cattle due to their ability of feature extraction and self-attention mechanisms, which enable precise classification even in complex visual scenes. YOLOv8 [4], an elite object detection model, is used for real-time cattle detection with high-speed and accurate localization of animals in different conditions. Opencows2020 dataset is used for Training and testing [5]. Dataset has 46 classes of Holstein Friesian cows. The dataset has sufficient diversified instances and it serves as a robust platform for experimentation. By integrating advanced machine learning models, this study aims to enhance the accuracy and efficiency of cattle detection and identification that ensures effective livestock management solutions.

Object Identification and detection are two different tasks of computer vision field. For efficient livestock monitoring focused is on cattle identification and detection. Both involve recognizing cattle in images or videos however, their objectives and methodologies differ.

### 1.1 Cattle Identification

The objective of cattle identification is precise detection and individual distinction of cattle from each other. It involves identifying the distinctive features of every animal (e.g., facial features, ear marks, muzzle types) for individuation [6]. CNN-based models such as FaceNet or bespoke CNN models are utilized to extract deep features from cattle faces to identify and separate individual animals [7][8]. Cattle face recognition systems based on CNNs are able to recognize cattle with high accuracy from their facial structure even when the animal is moving. Methods such as Siamese Networks are employed for cattle recognition by learning a metric space in which the similarity between images of the same cow is minimized and between different cows maximized. Deep metric learning

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

methods are often used for cattle ID tasks, where the system identifies individual cows based on biometric traits like ear tags or muzzle patterns. Models like BEiT (BERT Pretraining of Image Transformers) [9] use self-supervised learning for pre-training on large image datasets to learn rich representations, which are then fine-tuned for specific identification tasks. BEiT-based cattle ID systems can improve recognition by leveraging large-scale unlabeled datasets. Cattle identification requires high-quality data and is sensitive to variations in pose, occlusion, and appearance changes over time (e.g., from aging or changes in coat color).

### 1.2 Cattle Detection

The objective of cattle detection is to identify and classify the occurrence of cattle in an image or video frame. It generally entails the detection of the bounding boxes around the cattle, thus it is a localization problem [10]. Convolutional Neural Networks (CNNs) are commonly employed for cattle detection. Many authors uses models such as YOLO (You Only Look Once) and Faster R-CNN [11]. These models are trained to identify the cow as an object within an image and return its coordinates. YOLOv5 [12] is a popular model for cattle detection in open fields and it is capable of detecting cattle even images taken from long distance. YOLOv4 [13] and YOLOv3 [14] also have similar capabilities. DETR (Detection Transformer) [14] utilizes transformers for object detection by treating object localization as a set prediction problem. It eliminates the need for region proposals. DETR has been used to detect cattle in aerial imagery and it outperform traditional CNN-based detectors. In case of overlapping cattle, different lighting conditions, or partial occlusions detection shows suboptimal performance.

Cattle identification aims to uniquely recognize individual cattle in contrast to cattle detection which focuses on locating cattle within images. Deep learning and transformer models gives optimal result for both the tasks and achieve high accuracy and real-time processing capabilities. YOLO and DETR [14] are good at detection tasks, while CNNs, deep metric learning, and transformer-based methods like BEiT [9] and Siamese Networks are proved most optimal for cattle identification.

### 1.3 Motivation

Cattle recognition and detection are essential aspects of livestock management. The recognition and detection are depending on technologies and used for monitoring health, preventing theft, controlling disease, and allocating resources. Also, with the increasing demand for livestock products worldwide, there is a need for more effective and reliable cattle management practices. Cattle recognition techniques such as ear tagging, branding, or even RFID implants have some drawbacks which include discomfort to the animal, the tags being easily removed, or an expensive cost to implement them. There is a need, therefore, for automated methodologies that require little to no intervention during implementation using modern technologies. In this way, the use of computer vision systems for automated cattle recognition and tracking has emerged as a feasible solution through the processing of visual information. During the last years, the progress in deep learning applications led to a widespread use of convolutional neural networks (CNNs) for object detection and classification, including recognition of cattle. Recognition of objects using CNNs is often not very accurate due to occlusion, angle and position changes, illumination variation, and the different patterns of cattle fur. With animal image recognition, these issues have been remedied through more sophisticated network configurations such as Vision Transformers (ViT) and modern object detection methodologies.

### 1.4 Problem Statement

Researchers of computer vision realise the complexity to get precise and scalable cattle identification and detection system, even with advanced technology in deep learning and computer vision. Cattle detection requires a lot of time and effort using manual or even semi-automated techniques, yielding a lot of errors. Conventional computer vision techniques tend to fall short of the real-life challenges' thresholds which makes the task of cattle identification complicated. Another multilayered issue comes with variability within classes, similarity across classes, and including external factors such as light, weather, and obstruction to the surroundings inside a box. Moreover, the deployment of efficient and accurate identification systems in resource-limited farm environments is hindered by the possibilities of scale and perspective changes together with the increased computational cost of deep learning models.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

There is a growing demand to come up with models which are robust high-performance and can accurately detect and identify cattle under completely different real-world operating environments. Cattle detection and identification suffer from issues that can be tweaked with Vision Transformers and YOLOv8, which have proven successful in solving almost all problems.

### 1.5 Challenges in Cattle Identification and Detection

The automated detection and identification of cattle are beset with problems that need to be solved to allow for real-world application [15-23]. Cattle identification has a number of challenges that impact efficiency, accuracy, and scalability of animal husbandry. Traditional methods such as ear tags, branding, and RFID chips are subject to wear and tear, loss, and damage, leading to wrong tracking and misidentification. Manual observation is time-consuming, labor-intensive, tedious, and susceptible to human errors, particularly when working with large farming operations. Additionally, variations in cattle appearances, including coat color patterns, age, and weather conditions like light, make it even more difficult through visual identification. Sustainable real-time monitoring in dynamic farm environments makes system more complex and it need robust and effective solutions that can cope with uncontrolled environments. Further, most of the conventional systems are not mechanized and rely on human observations, which makes the system ineffective and costly in the long run. Due to this, the installation of advanced, technology-driven solutions that are able to provide accurate, scalable, and real-time cattle identification and detection is needed.

Cow detection has many challenges that need to be addressed for effective and efficient monitoring of livestock. One of the main challenges is the variation in the appearance which make recognition inconsistent and difficult. Variation arise due to breed, size, posture, coat patterns, or any combination thereof. The vision systems become more complex due to environmental factors such as inadequate lighting, occlusions, shadows, or background clutter, especially in open fields, or in crowded farms. In addition, real time detection is only possible when the systems are optimized for high speed processing and can operate effectively on edge devices with low computational capacity. Motion blur along with partial occlusions where some cattle are blocked by other objects or animals makes it difficult for conventional vision systems to track accurately. Identifying individual cattle in a group is another major hurdle because overlapping bodies can result in misclassification or missing data altogether. Furthermore, most of the current systems for detection do not work in active farm settings where there is a lot of movement, re-positioning, or interaction among the cows. Meeting these objectives entails the use of sophisticated deep learning approaches and enhancement of the detection’s speed, accuracy, and reliability in real agricultural field conditions.

## OBJECTIVES

The overall aim of this work is to create cattle identification and detection system based on Vision Transformers and YOLOv8 [4] for on the Opencows2020 dataset [5]. Our target is to design a strong system for correctly detecting and identifying unique cattle in diverse real-world cases. We seek to compare the performance of Vision Transformers and YOLOv8 in accuracy, efficiency, and generalization capacity while tackling main challenges like occlusion, moving backgrounds, and pose variations. We also target optimizing these models for real-time execution and real-world deployment within resource-limited environments to enhance their practical applicability in agricultural environments. This paper seeks to enhance automated cattle identification and detection using the strengths of Vision Transformers and YOLOv8, paving the way for more efficient and scalable approaches in modern livestock management.

In this paper, the related work in cattle detection and identification is reviewed in Section 2. The suggested methodology, including the YOLOv8 and Vision Transformers structures, is described in Section 3. While Section 5 addresses the study's evaluation metrics, Section 4 presents the OpenCows2020 dataset. Following the presentation of the experimental setup and results in Section 6, the paper's main conclusions and insights are provided in Section 7.

## LITERATURE REVIEW

The literature discusses various techniques that are effectively utilized for cattle identification, emphasizing the importance of accurate identification. Mainly these researches addresses application areas such as genetic

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

improvement, disease control, biosecurity, and efficient supply chain management. Farmers have implemented personalized management strategies, recognizing that effective cattle identification is a critical factor in achieving optimize results. Traditional identification methods are ear tagging, DNA analysis, and visual feature-based approaches. Biometric indicators for cattle identification include DNA profiling, antibody matching, muzzle pattern recognition, and facial feature analysis [11]. These biometric methods leverage both phenotypic and genomic characteristics that are unique to each animal. They are resistant to tampering, stable over time, and cause minimal impact on the animal's health, making them reliable and efficient for identification purposes.

The literature review mainly focus on machine learning (ML) and deep learning (DL) techniques with image processing and computer vision for cattle identification. The review looks at several aspects of cattle identification techniques and groups the researches and techniques to provide an extensive overview of the area. Among the categories that have been identified are:

- Hardware Aspect: Approaches that rely on physical tags for identification.
- DNA Aspect: Methods based on DNA features for identification purposes.
- Biometric Aspect: It utilize visual features for cattle identification.
- Machine Learning Aspect: Method using machine learning algorithms for identification.
- Deep Learning Aspect: Approaches that leverage deep learning models for enhanced identification accuracy.
- Integrated Learning Aspect: Combined ML and DL approaches that integrate both techniques for improved cattle identification.

These categories include the various and evolving approaches being investigated in the field of cattle identification, showing the growing importance of modern technology in improving the accuracy, efficiency, and scalability of cattle management systems. The combination of classic and cutting-edge techniques, specifically machine learning and deep learning, has the potential to greatly improve cattle identification systems.

The development of deep learning and transformer structures has greatly improved cattle identification methods. Muzzle Based Identification CNN Transformer fusion method [24] captures both local and global features using Convolutional Neural Networks (CNNs) and transformer models from cattle muzzle images. The fusion of these architectures overcomes the limitations of CNNs for long-range dependencies, thereby increasing identification accuracy. Unified Deep Learning Framework Using Video for Analysis, A multi CNN BiDirection Long Short Term Memory (BiLSTM) and self-attention video data processor trained with a sophisticated technique. This framework captures spatio-temporal features to accurately identify individual cattle in active surroundings. Multi Feature Decision Level Fusion technique [26] uses face, muzzle, and ear tag features of cattle to make accurate identification at the decision level fusion. The combination of various biometric traits resulted in robust and reliable identification. Parallel Attention Network for Cattle Face Recognition [27] implements a novel design using modules in parallel attention in a transformer that focuses on both local and global features of cattle faces. This method improves recognition accuracy, particularly in challenging wild environments. YOLOv5 with Transformer for Muzzle Pattern Detection [28] is a method using the YOLOv5 object detection model incorporating transformer modules to identify cattle using muzzle patterns. The inclusion of transformers assists in capturing intricate patterns, thus providing high identification accuracy. Such methods illustrate the power of unifying deep learning and transformer architecture in streamlined cattle identification systems.

## TRANSFORMERS: VIT, DEIT, AND BEIT AND YOLOV8

Vision Transformers (ViTs) have transformed computer vision by presenting a radically new way of feature extraction. Contrary to hierarchical feature maps-based conventional CNNs, ViTs utilize self-attention to learn global dependencies in images. This makes the model more optimal in fine-grained recognition tasks like cattle identification. Additionally, the capacity of ViTs to deal with occlusion, pose changes, and intricate patterns qualifies them for this field. However, the object detection models YOLOv8, balances state-of-the-art performance with real-time speed. YOLOv8 brings improvement in feature extraction, bounding box regression, and confidence scoring that makes it highly efficient in the detection of cattle in adverse environments. Its light-weight design

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

makes it possible to deploy it on edge devices, which is perfect for practical farm implementation. Transformers and YOLOv8 are elaborated as follows:

### 4.1 ViT

The Vision Transformer (ViT) proposed by Dosovitskiy et al. [1] in 2020 transformed computer vision by using transformer architectures. It was initially developed for natural language processing and for image recognition tasks. ViT works by splitting the images into fixed-size patches (for example, 16x16 pixels), flattening them, and embedding them into a vector space and treating each of them as a token of sequence. This sequence is then passed through a regular transformer encoder, which applies self-attention mechanisms to capture global dependencies in the image. A learnable classification token is prepended to the sequence, and its output serves for image classification. One of the largest benefits of ViT is that it is capable of capturing global context in images, which results in better performance on tasks involving holistic understanding. The ViT scales well and generates competitive performance than conventional CNNs when pre-trained on large datasets. However, ViTs suffer from issues like data inefficiency, needing large datasets for efficient training, and high computational requirements for large models. These disadvantage is because of its lack of CNN-like inductive biases. Recent studies focuses on these limitations using hybrid models that incorporate CNNs and ViTs. The resultant architecture is computationally efficient with reduced computational requirements, and investigations of the robustness and generalization potential of ViTs. Dosovitskiy et al. (2020) [1], Papa et al. (2023) [29], and Naseer et al. (2021) [30], provide in-depth review of the design, performance, and innovations of Vision Transformers.

### 4.2 DeiT

Data-efficient Image Transformer (DeiT) proposed by Touvron et al. in 2020 [2], uses a novel distillation process to improves the Vision Transformer (ViT) architecture in terms of better data efficiency. DeiT is different from other distillation techniques as it uses an exclusive distillation token that is engaged with the class token within the attention process of the transformer. This architecture learn well from a teacher network, such as a convolutional neural network (CNN), thus introducing CNN-like inductive biases without using convolutional layers. Consequently, without the need for large external data, the DeiT acquires competitive performance on image classification only with the ImageNet dataset. This finding shows that transformers are efficiently trainable for vision tasks by utilizing novel distillation methods even with a small amount of data.

### 4.3 BeiT

BEiT - Bidirectional encoder representation from image transformers, is a supervised vision model proposed by Hangbo et. al. in 2021 [3]. Inspired by BERT, a known model in the field of natural language processing, BEiT adopts a masked image modeling (MIM) strategy to pre-train vision transformers. In this model, an image is split into patches (say, 16×16 pixels) and tokenized into discrete visual tokens. In pre-training, some patches are randomly masked and the model is trained to recover the original visual tokens from corrupted inputs. This approach allows the model to learn subtle visual representations without using pre-labelled data. Experimental results have shown that BEiT has competitive performance in image classification and semantic segmentation and outperforms conventional models.

### 4.4 YOLOv8

YOLOv8 is object detection algorithms and it gives noticeable architectural advancements in terms of performance and resilience [4]. Some major advances include the addition of a Cross Stage Partial Network (CSPNet) backbone for improved feature extraction and an integrated Feature Pyramid Network (FPN) and Path Aggregation Network (PAN) neck for enhanced multi-scale object detection. The YOLOv8 shifts to an anchor-free design which streamline the detection process and keeps computational complexity low. YOLOv8 mange obtain state-of-the-art performance in object detection with real-time processing, which is much needed for applications such as autonomous vehicles and video surveillance. Yaseen [31] and Varghese et. al. [4] offered extensive examination of YOLOv8's architecture and performance gains and observations on its design and effectiveness.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

## OPENCOW2020 DATASET

Opencows2020 dataset contains images of 46 classes of cattle. The OpenCow2020 dataset [32][5] is a very large dataset containing images to allow the detection, localization, and identification of unique Holstein Friesian cows using deep learning methods. Built by researchers from the University of Bristol, this dataset contains 11,779 images with 13,026 labelled objects belonging to the class of cows. The scenes contain indoor and outdoor top-view images, allowing for varied test scenarios for model training and assessment. One key aspect of OpenCow2020 is its capability for open-set identification, where cattle not seen during training are recognized without system retraining. This is very useful for actual applications in precision agriculture, as herds in the field change dynamically. The data is carefully split into training, validation, and test subsets to enable strong cross-validation and model evaluation. In their supporting research, the authors utilized convolutional neural networks and deep metric learning methods and achieved a cattle identification accuracy of 93.8% when trained on a half-sample of the cattle population. This highlights the dataset's potential for supporting the development of non-intrusive livestock monitoring systems.

## EVALUATION METRICS

Cattle identification is critical in livestock management, traceability, disease control, and food safety. Automated identification systems depend on some key performance parameters like accuracy, identification rate, durability, scalability, ease of use, security, and privacy to ensure they remain efficient [11]. Accuracy determines how effectively the system identifies individual cattle, impacting health monitoring, breeding, and regulatory compliance. The identification rate measures correct identifications, while error rates, including false positives and false negatives, highlight misclassifications. Speed is crucial for large herds, enabling real-time tracking, while durability ensures hardware reliability in harsh conditions. Scalability allows the system to adapt to herd expansion, and integration with farm management tools enhances interoperability with health, feeding, and breeding systems. For performance evaluation confusion matrix is form that contains true positive, true negative, false positive and false negative. From confusion matrix we calculate accuracy, true positive rate, false positive rate, true negative rate, and false negative rate. Precision, recall, and the F1-score are used to provide holistic system effectiveness [11]. Continuous real-time monitoring through field testing, data logging, and long-term durability assessments ensures reliable performance over the period. It help farms in optimize cattle identification for efficient herd management and desire regulatory compliance.

$$True\ Positive\ Rate\ = \frac{Correct\ Identifications}{Total\ Identifications} * 100$$

$$False\ Positive\ Rate = \frac{False\ Identifications}{Total\ Identifications} * 100$$

$$True\ Negative\ Rate = \frac{Correct\ Exclusions\ (True\ Negatives)}{Total\ Non - Target\ Cattle} * 100$$

$$False\ Negative\ Rate = \frac{Missed\ Identifications}{Total\ Identifications} * 100$$

Macro-average and weighted-average, precision, recall, and F1 scores were computed, alongside overall accuracy [33]. Mean Average Precision (mAP) evaluate the model across multiple classes in detecting and localizing objects for detection. Precision gives the fraction of correctly detected instances among the retrieved instances. Higher precision means fewer false positives. Recall is the fraction of relevant instances that have been retrieved. Higher recall means fewer false negatives. F1-Score is a harmonic mean of precision and recall, useful when the dataset is imbalanced. Top-1 and Top-5 Accuracy is often used for classification tasks (e.g., cattle identification). Top-1 measures the accuracy of the single predicted label, while Top-5 measures whether the correct label is among the top 5 predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

$$Precision = \frac{TP}{TP + FP} * 100$$

$$Recall = \frac{TP}{TP + FN} * 100$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recal} * 100$$

## METHODOLOGY AND RESULTS

Cattle Identification and Detection Approaches, Methodologies and associated results on Opencows2020 dataset are discussed in this section.

### 7.1 Cattle Identification Methodology

Vision transformers (ViT, DeiT, and BEiT) were modelled on the Opencows2020 dataset to predict cattle according to coat patterns. Training is done by PyTorch in conjunction with Hugging Face Transformers, with a batch size of 64 for effective processing. The AdamW optimizer is used with 0.05 weight decay for increased generalization, and the learning rate is used as 2e-4, which is then cosine annealed for smooth convergence. Training is done for 05 epochs with the objective function being cross-entropy loss to distinguish among cattle coat patterns. To support faster computations and handle big data, NVIDIA A100 GPUs are employed in a multi-GPU environment to allow for faster training and more scalability. The pre-processing pipeline maximizes the OpenCows2020 dataset to support training of Vision Transformers (ViT, DeiT, and BEiT) for cattle coat pattern classification.

To facilitate generalization, several data augmentation strategies including random cropping, flipping, rotation, and contrast modifications are utilized, mimicking real-world variations such as occlusions and lighting variations. Normalization is achieved by normalizing pixel values to maintain uniformity across images, whereas resizing is important to achieve compatibility with various transformer models—224×224 for ViT and DeiT, and 384×384 for BEiT. Besides, BEiT needs tokenization to occur, wherein the images are encoded into discrete patch embedding to conduct masked image modelling, enhancing the learning of features.

The model structures utilize the self-attention mechanism to accurately classify cattle by coat pattern. ViT initially splits an image into 16×16 patches, flattens them, and projects them into an embedding space, followed by positional encoding to preserve spatial relationships. The embeddings go through multi-head self-attention (MSA) layers, a feed-forward network (FFN), and layer normalization, finally employing a classification token (CLS token) for prediction. DeiT follows a similar structure but introduces a distillation token, allowing the model to learn from a CNN-based teacher, making it more data-efficient and enhancing training with techniques like RandAugment, MixUp, and CutMix. Meanwhile, BEiT is based on self-supervised pretraining, employing a masked image modelling (MIM) strategy, where portions of the image are randomly masked and reconstructed, akin to BERT in NLP. This helps the model learn contextual and structural details crucial for distinguishing coat patterns. All models incorporate classification heads that map the learned representations to distinct cattle coat categories, ensuring effective cattle identification.

### 7.1.1 Cattle Identification Result

The results indicate that the DeiT Transformer achieves the highest accuracy (99.80%), closely followed by the ViT Transformer (99.79%), both demonstrating exceptional performance across all metrics. They exhibit near-perfect macro and weighted average precision, recall, and F1 scores, highlighting their ability to maintain consistency across classes. ResNet-50, with an accuracy of 98.0%, performs slightly lower but still achieves competitive precision, recall, and F1 scores, showcasing its robustness. In contrast, the BEiT Transformer lags significantly behind with an accuracy of 95.97% and notably lower macro average metrics (0.913), suggesting challenges in balancing performance across all classes. This analysis highlights the superior capabilities of transformer-based architectures, particularly DeiT and ViT, for plant disease detection tasks, with ResNet-50 as a strong convolutional baseline and BEiT requiring improvements for enhanced performance.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Table 1:** Cattle Identification Results on Various Transformers compared to ResNet-50

| Model | Accuracy | Macro Avg Precision | Weighted Avg Precision | Macro Avg Recall | Weighted Avg Recall | Macro Avg F1 Score | Weighted Avg F1 Score |
|---|---|---|---|---|---|---|---|
| ResNet-50 | 98.0% | 0.970 | 0.984 | 0.980 | 0.982 | 0.973 | 0.982 |
| ViT Transformer | 99.79% | 0.999 | 0.998 | 0.996 | 0.998 | 0.997 | 0.998 |
| DeiT Transformer | 99.80% | 0.999 | 0.998 | 0.996 | 0.998 | 0.997 | 0.998 |
| BEiT Transformer | 95.97% | 0.913 | 0.969 | 0.913 | 0.969 | 0.913 | 0.969 |

The Figure 1 appears to showcase results from a cattle identification model using BEiT Transformers. Each cell in the grid displays an image of cattle, annotated with the true label and the predicted label, reflecting the model's performance.

Most cells in the grid indicate that the BEiT Transformer has correctly identified the cattle, as the predicted labels match the true labels. For example, rows with "True Label: 2, Predicted Label: 2" or "True Label: 1, Predicted Label: 1" demonstrate accurate identification. A few instances of misclassification are observed. For example, one cell indicates "True Label: 35, Predicted Label: 6." These misclassifications highlight potential limitations of the model in distinguishing between certain cattle, possibly due to similar patterns or features among individuals.

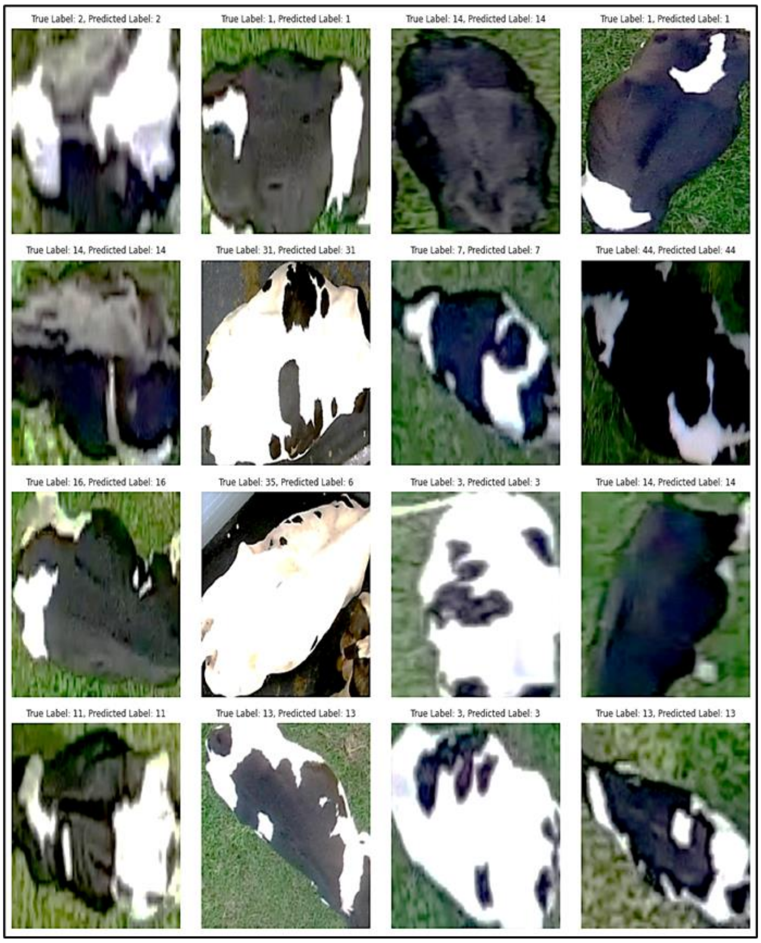
True Label: 2, Predicted Label: 2
True Label: 1, Predicted Label: 1
True Label: 14, Predicted Label: 14
True Label: 1, Predicted Label: 1
True Label: 14, Predicted Label: 14
True Label: 31, Predicted Label: 31
True Label: 7, Predicted Label: 7
True Label: 44, Predicted Label: 44
True Label: 16, Predicted Label: 16
True Label: 35, Predicted Label: 6
True Label: 3, Predicted Label: 3
True Label: 14, Predicted Label: 14
True Label: 11, Predicted Label: 11
True Label: 13, Predicted Label: 13
True Label: 3, Predicted Label: 3
True Label: 13, Predicted Label: 13

**Figure 1:** Cattle Identification on Opencows2020 dataset using BeIT Transformers

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Some images show cattle in varying poses or partially occluded, which may complicate feature extraction and recognition. However, the model still performs well in many of these challenging cases. Several images appear blurred or lack fine details, which can affect the model's ability to extract distinguishing features. The BEiT model's performance in such cases is mixed, as seen from the accurate and misclassified examples. Certain cattle with highly similar fur patterns (e.g., black-and-white patches) may lead to confusion, as seen in the incorrect classifications. This suggests that BEiT Transformers might struggle with fine-grained feature discrimination when cattle have overlapping visual characteristics.

BEiT Transformers effectively handle a variety of poses, scales, and occlusions in many instances, demonstrating their robustness in extracting global features using self-attention mechanisms. The misclassifications reflect issues in resolving minor intra-class variation or dealing with low-quality cases or ambiguous patterns. Additional training on a large dataset and involving methods such as data augmentation or fine-grained feature learning could enhance the BEiT model performance for cattle classification.

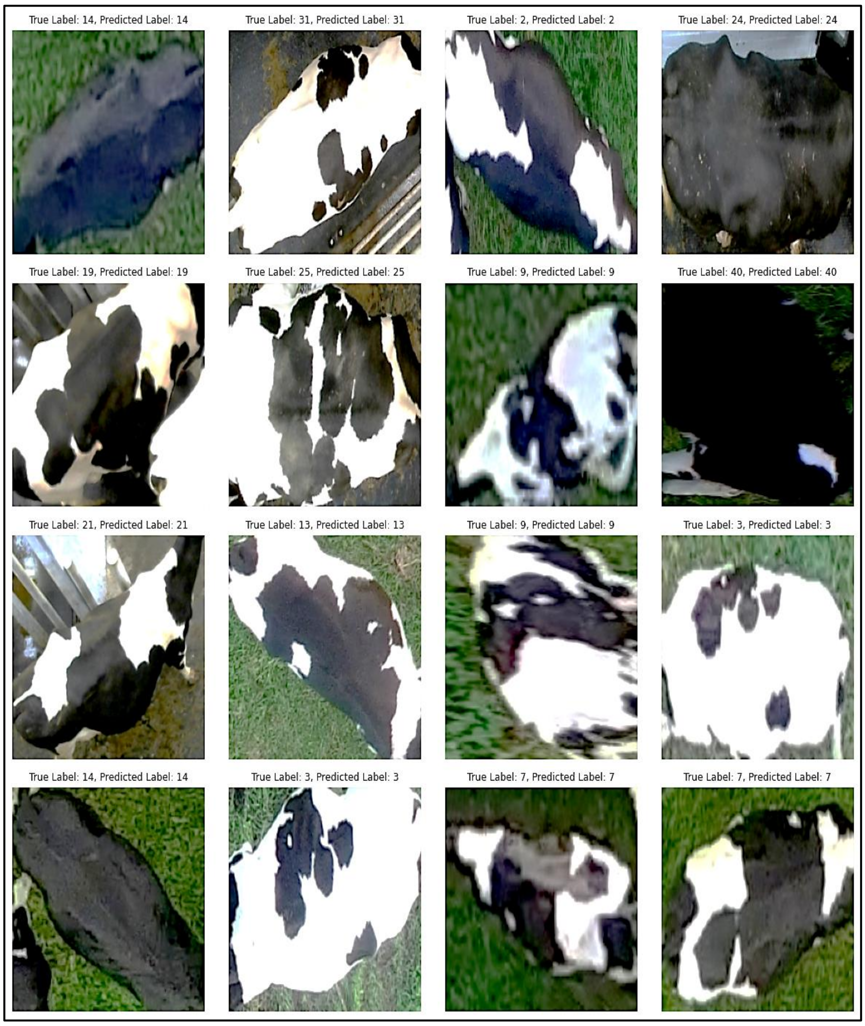
True Label: 14, Predicted Label: 14
True Label: 31, Predicted Label: 31
True Label: 2, Predicted Label: 2
True Label: 24, Predicted Label: 24
True Label: 19, Predicted Label: 19
True Label: 25, Predicted Label: 25
True Label: 9, Predicted Label: 9
True Label: 40, Predicted Label: 40
True Label: 21, Predicted Label: 21
True Label: 13, Predicted Label: 13
True Label: 9, Predicted Label: 9
True Label: 3, Predicted Label: 3
True Label: 14, Predicted Label: 14
True Label: 3, Predicted Label: 3
True Label: 7, Predicted Label: 7
True Label: 7, Predicted Label: 7

**Figure 2:** Result of Cattle Identification on Opencows2020 dataset using DeIT Transformers

The pictures depicted in Figure 2, shows the result of a cattle identification model from DeiT (Data-efficient Image Transformer) Transformers. Most of the examples demonstrate properly matched true and predicted labels, like "True Label: 14, Predicted Label: 14" or "True Label: 31, Predicted Label: 31." This demonstrates the DeiT's correct identification of individual cattle in most cases. In given dataset, the images contain cattle in various orientations and poses. Still the model perform optimally in such instances which indicates its capacity to work with varied angles and views. Some images contain partial occlusion or shadows that hide important features. However, the DeiT transformer is able to achieve high accuracy in such situations. Result shows less misclassifications in this

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

particular sample which suggest that the DeiT model is very strong. Still, performance needs to be assessed on large dataset to validate generalization and stability.

The model seems to differentiate well between slight variations in fur patterns and markings among cattle, which is vital for correct identification. In the presence of diversified and noisy environments, e.g., grass and agricultural equipment, the model is shown to be robust in the detection and identification of cattle. DeiT Transformers utilize self-attention and efficiency of data to enhance performance where fine-grained recognition and feature extraction is needed. This shows the model's ability to differentiate appropriately between cattle under complicated and dynamic real-world scenarios. To address edge cases such as extreme occlusion or low illumination, data augmentation, transfer learning, or fine-tuning would be useful. The outcomes show that DeiT Transformers are more optimal for cow detection which indicates their ability to manage real-world complexity with great accuracy.

The images in Figure 3 indicate the output of a Vision Transformer-based cattle identification model. Every grid cell presents a cattle image along with its actual and predicted labels, which convey details about the capability of the model to classify. Most of the samples in the grid exhibit correct predictions, as the actual labels are similar to the expected labels. For instance, instances like "True Label: 9, Predicted Label: 9" and "True Label: 17, Predicted Label: 17" reflect proper classification. The images contain cattle in multiple orientations and attitudes. The ViT model identifies cattle in most environments properly, reflecting that its self-attention mechanism achieves unique features irrespective of orientation. Even with variations in background contexts, e.g., grass, soil, or barn floors, the ViT model's performance is not affected, showing resilience to context changes.

The ViT model does a great job in distinguishing subtle patterns, such as can be observed from images where correctly identified cattle of black-and-white spots are involved. This draws attention to the strength of the model in capturing global context through self-attention mechanisms. Several images are darker or occluded partially. Owing to this adversity, still the model scores accurately in the majority of the cases, i.e., "True Label: 44, Predicted Label: 44". Although not immediately apparent in this particular sample, there may be misclassifications that occur in edge cases where two cattle have very similar patterns or when dominant features are occluded. The data set seems to be varied, with cattle differing in color, size, and pattern. The model's correct classification of these varied examples speaks to the ViT's strength in dealing with real-world nuance.

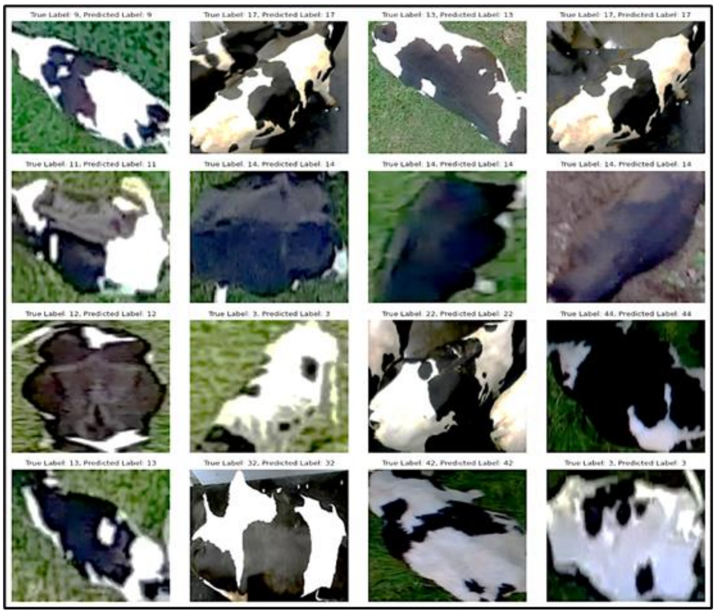
True Label: 9, Predicted Label: 9
True Label: 17, Predicted Label: 17
True Label: 13, Predicted Label: 13
True Label: 17, Predicted Label: 17
True Label: 11, Predicted Label: 11
True Label: 14, Predicted Label: 14
True Label: 14, Predicted Label: 14
True Label: 14, Predicted Label: 14
True Label: 12, Predicted Label: 12
True Label: 3, Predicted Label: 3
True Label: 22, Predicted Label: 22
True Label: 44, Predicted Label: 44
True Label: 13, Predicted Label: 13
True Label: 32, Predicted Label: 32
True Label: 42, Predicted Label: 42
True Label: 3, Predicted Label: 3

**Figure 3:** Results of Cattle Identification on Opencows2020 dataset using ViT Transformers

### 7.2 Cattle Identification Methodology

The Opencows2020 dataset contained images of cattle with varying coat patterns, lighting, and angles of view. To make dataset suitable for training and to optimize performance, the dataset rigorously pre-processed and augmented. The dataset pre-process and transformed into a YOLO-compatible format with labelled bounding

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

boxes. Pre-processing involved resizing images to standard sizes, normalizing pixel values for uniformity, and transforming annotations into text files for smooth model training. To ensure generalise model, diverse data augmentation processes were used which includes random resizing, color jittering, addition of Gaussian noise, horizontal flip, rotation, and Mosaic augmentation. In mosaic augmentation, multiple images were combined to enhance detection of small objects. Such transformations helped to train the model to recognize cattle under different conditions, enhancing robustness of model.

The YOLOv8 model architecture [4][31] is an extension of the earlier YOLO models. It improves detection precision and decrease processing time. The architecture includes three key parts: the Backbone, Neck, and Head. The Backbone is based on CSPDarkNet53. It integrates Cross-Stage Partial Networks (CSPNet) for extracting features and Efficient Layer Aggregation Networks (ELAN) for enhancing gradient flow with less computation requirement. To enhance spatial feature learning, squeeze-and-excitation (SE) blocks are also used. Neck employs a Path Aggregation Network (PAN) to fine-tune multi-scale feature maps so that accurate cattle detection can be achieved under different conditions. Head adopts anchor-free detection, decoupled heads with individual classification and regression, and Dynamic Label Assignment (DLA) for better generalization. With its optimized and lightweight architecture, YOLOv8 is capable of high accuracy in coat pattern recognition of cattle and can keep up real-time processing, rendering it highly suitable for actual livestock management purposes.

The model was trained using PyTorch and the YOLOv8 framework with batch sizes ranging from 32 to 64, considering availability of GPU memory. The AdamW optimizer was used with a weight decay of 0.01 to improve stability and to prevent over-fitting. A learning rate of 1e-3 was used, along with a cosine annealing scheduler for improved convergence. The model was trained on NVIDIA A100 GPUs using PyTorch's Distributed Data Parallel (DDP) approach, which allows for efficient multi-GPU training while maintaining optimal performance and scalability. The model was trained for 300 epochs, reducing three critical loss components (as shown in Table 1): box loss, classification loss, and distributional focal loss (DFL loss), all of which decreased significantly which indicates improved accuracy.

**Table 2:** Performance analysis using loss metrics for cattle detection using YOLOv8 on Opencows2020 dataset

| Loss Metrics | Initial Value | Final Value | Key Observations |
|---|---|---|---|
| **Box Loss** | 3.29 | 1.05 | Rapid decline in first 50 epochs, indicating improved bounding box accuracy. |
| **Classification Loss** | 3.63 | 0.65 | Significant drop after 80 epochs, reflecting better class differentiation. |
| **Distributional Focal Loss (DFL Loss)** | 4.11 | 1.5 | Gradual decline, showing improved bounding box quality. |

### 7.2.1 Cattle Detection Results

Evaluation measures shown in Table 3 exhibits optimal performance, with accuracy of 0.993, recall of 0.98, and mAP@50-95 of 0.94. This suggests the model's reliability in identifying cattle with precision. Low over-fitting and consistent validation trends also indicates YOLOv8's applicability in real-world agricultural scenarios. This makes the system dependable for automated livestock monitoring.

**Table 3:** Evaluation Metrics for detection Result

| Evaluation Metrics | Value | Key Observations |
|---|---|---|
| **Precision** | 0.993 | Reached 0.99 by epoch 40, stable afterward, indicating a low false positive rate. |

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

| | | |
|---|---|---|
| **Recall** | 0.98 | Improved significantly in the first 50 epochs, enhancing detection capability. |
| **mAP@50** | 0.99 | High precision achieved across different IoU thresholds. |
| **mAP@50-95** | 0.94 | Reached stability after epoch 80, indicating robustness across IoU thresholds. |

Training and validation performance metrics of the model over 300 epochs are illustrated in Figure 4. Distributional focal loss (DFL), box, and classification loss curves indicate steep drops within the first 50 epochs followed by stabilization, which reflects proper optimization. There is no significant over-fitting as trends in validation loss are nearly as good as in training loss. Early training brings remarkable improvements in indicators like accuracy, recall, and mean average precision (mAP), where the recall hits 0.98 and the precision hits 0.993. The excellent detection performance over all IoU threshold values is upheld by high mAP@50 (0.99) and mAP@50-95 (0.94). This demonstrates very impressively how the YOLOv8 recognizes cattle with an excellent recall at minimal false alarms.

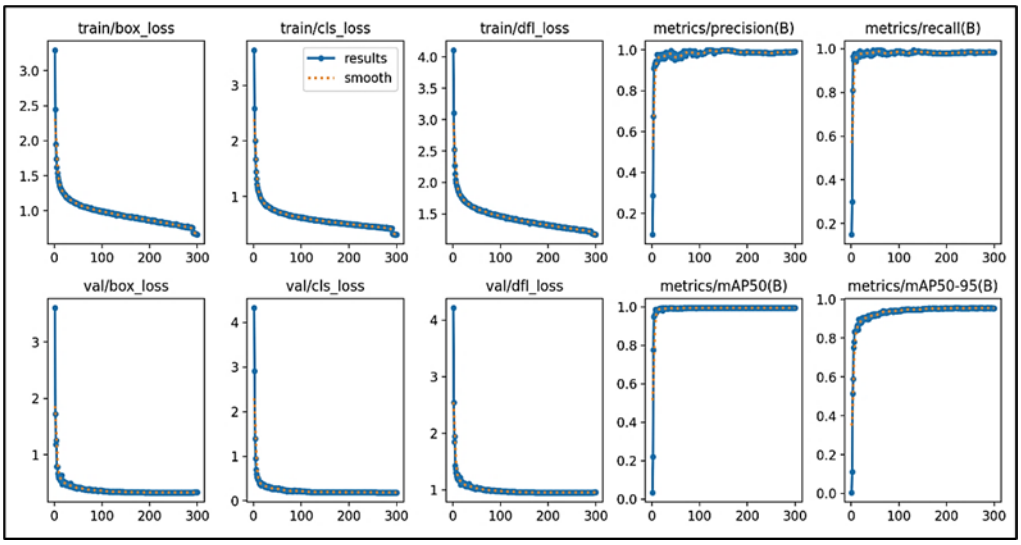

**Figure 4:** Training and Validation performance comparison for Cattle Detection Model

The Figure 5 illustrates the YOLOv8 model's performance in detecting cattle, with red bounding boxes surrounding identified cows in different environments. The model is able to detect cattle from aerial perspectives, indoor environments, and cluttered backgrounds, proving its versatility. Different coat patterns, lighting conditions, and orientations are well recognized, indicating the robustness of the model. However, some overlapping or slightly misaligned bounding boxes suggest minor detection inconsistencies, though overall accuracy appears high. The results confirm that the model can generalize across diverse scenarios, making it suitable for real-world cattle monitoring applications.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

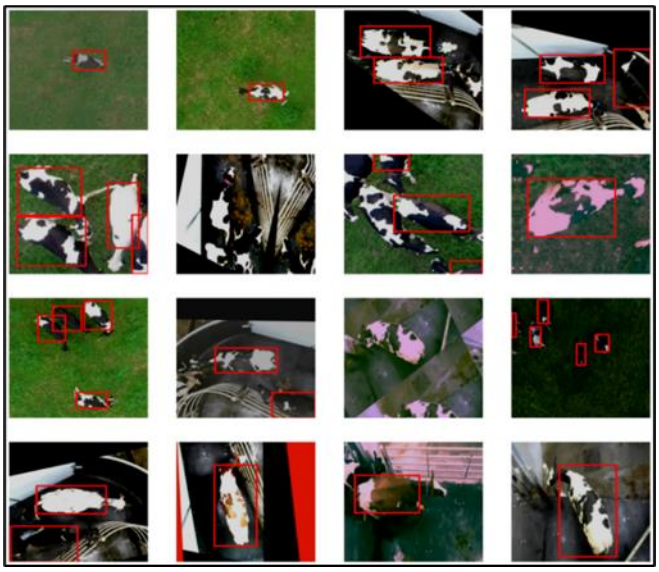

**Figure 5:** Cattle Detection using YOLOv8 model

The ViT model's self-attention mechanism is very good at capturing global features, which makes it suitable for difficult tasks like cattle identification. Background noise and pose variation robustness indicate very good generalizability. Misclassifications in more difficult cases (not depicted here) may indicate the need for additional fine-tuning or other forms of data augmentation to handle extreme pose variation or lighting. Incorporation of multi-scale feature extraction would further enhance performance in situations of minor differences in patterns. Vision Transformers also possess similarly good accuracy and robustness in cattle identification, and they can process the complexity and variability of real-world data.

## CONCLUSION

This research showcases the capability of deep learning methods in realisation of improved cattle management systems. Vision transformers are superior to conventional convolutional networks for detection, while YOLOv8 provides effective detection performance. The findings confirm the effectiveness of vision transformers, especially ViT and DeiT, in realizing optimal accuracy for the cattle identification task. The performance of YOLOv8 confirms its applicability to real-time detection. These results carry important implications in terms of promoting livestock management improvement, minimizing the need for human intervention, and providing scalability on large herd sets. Future work will proceed with embedding these models within actual applications such as mobile-based frameworks and smart farm solutions.

## REFRENCES

[1] Alexey, Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv: 2010.11929 (2020).

[2] Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." Advances in Neural Information Processing Systems 34 (2021): 23296-23308.

[3] Papa, Lorenzo, Paolo Russo, Irene Amerini, and Luping Zhou. "A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[4] Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." Advances in Neural Information Processing Systems 34 (2021): 23296-23308.

[5] Touvron, Hugo, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. "Training data-efficient image transformers & distillation through attention." In International conference on machine learning, pp. 10347-10357. PMLR, 2021.

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

[6] Awad, Ali Ismail. "From classical methods to animal biometrics: A review on cattle identification and tracking." Computers and Electronics in Agriculture 123 (2016): 423-435.

[7] 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

[8] Kusakunniran, W., Phongluelert, K., Sirisangpaival, C., Narayan, O., Thongkanchorn, K., & Wiratsudakul, A. (2023, October). Cattle AutoID: Biometric for Cattle Identification: Cattle AutoID. In Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology (pp. 570-574).

[9] Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

[10] Barbedo, J. G. A., Koenigkan, L. V., Santos, T. T., & Santos, P. M. (2019). A study on the detection of cattle in UAV images using deep learning. Sensors, 19(24), 5436.

[11] Hossain, M. E., Kabir, M. A., Zheng, L., Swain, D. L., McGrath, S., & Medway, J. (2022). A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture, 6, 138-155.

[12] Wang, W., Xie, M., Jiang, C., Zheng, Z., & Bian, H. (2024, June). Cow Detection Model Based on Improved YOLOv5. In 2024 39th Youth Academic Annual Conference of Chinese Association of Automation (YAC) (pp. 1697-1701). IEEE.

[13] Petso, T., Jamisola, R. S., Mpoeleng, D., & Mmereki, W. (2021, October). Individual animal and herd identification using custom YOLO v3 and v4 with images taken from a uav camera at different altitudes. In 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP) (pp. 33-39). IEEE.

[14] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

[15] Yao, Liyao, Zexi Hu, Caixing Liu, Hanxing Liu, Yingjie Kuang, and Yuefang Gao. "Cow face detection and recognition based on automatic feature extraction algorithm." In Proceedings of the ACM turing celebration conference-china, pp. 1-5. 2019.

[16] Yang, Zehao, Hao Xiong, Xiaolang Chen, Hanxing Liu, Yingjie Kuang, and Yuefang Gao. "Dairy cow tiny face recognition based on convolutional neural networks." In Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, China, October 12–13, 2019, Proceedings 14, pp. 216-222. Springer International Publishing, 2019.

[17] Kumar, Santosh, Sanjay Kumar Singh, and Amit Kumar Singh. "Muzzle point pattern based techniques for individual cattle identification." IET Image Processing 11, no. 10 (2017): 805-814.

[18] Kumar, Santosh, Amit Pandey, K. Sai Ram Satwik, Sunil Kumar, Sanjay Kumar Singh, Amit Kumar Singh, and Anand Mohan. "Deep learning framework for recognition of cattle using muzzle point image pattern." Measurement 116 (2018): 1-17.

[19] Chen, Shunnan, Sen Wang, Xinxin Zuo, and Ruigang Yang. "Angus cattle recognition using deep learning." In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 4169-4175. IEEE, 2021.

[20] Alzubaidi, Laith, Jinglan Zhang, Amjad J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions." Journal of big Data 8 (2021): 1-74.

[21] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15, no. 1 (2014): 1929-1958.

[22] Qiao, Yongliang, Daobilige Su, He Kong, Salah Sukkarieh, Sabrina Lomax, and Cameron Clark. "BiLSTM-based individual cattle identification for automated precision livestock farming." In 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), pp. 967-972. IEEE, 2020.

[23] Andrew, William, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. "Visual identification of individual Holstein-Friesian cattle via deep metric learning." Computers and Electronics in Agriculture 185 (2021): 106133.

[24] Dulal, Rabin, Lihong Zheng, and Muhammad Ashad Kabir. "MHAFF: Multi-Head Attention Feature Fusion of CNN and Transformer for Cattle Identification." *arXiv preprint arXiv:2501.05209* (2025).

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

[25] Qiao, Yongliang, Cameron Clark, Sabrina Lomax, He Kong, Daobilige Su, and Salah Sukkarieh. "Automated individual cattle identification using video data: a unified deep learning architecture approach." Frontiers in Animal Science 2 (2021): 759147.

[26] 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

[27] Li, Jiayu, Xuechao Zou, Shiying Wang, Ben Chen, Junliang Xing, and Pin Tao. "A Parallel Attention Network for Cattle Face Recognition." arXiv preprint arXiv:2403.19980 (2024).

[28] Dulal, Rabin, Lihong Zheng, Muhammad Ashad Kabir, Shawn McGrath, Jonathan Medway, Dave Swain, and Will Swain. "Automatic cattle identification using yolov5 and mosaic augmentation: A comparative analysis." In 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), pp. 1-8. IEEE, 2022.

[29] Papa, Lorenzo, Paolo Russo, Irene Amerini, and Luping Zhou. "A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking." IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[30] Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." Advances in Neural Information Processing Systems 34 (2021): 23296-23308.

[31] Yaseen, M. "What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector. arXiv 2024." arXiv preprint arXiv:2408.15857.

[32] Andrew, William, Jing Gao, Siobhan Mullan, Neill Campbell, Andrew W. Dowsey, and Tilo Burghardt. "Visual identification of individual Holstein-Friesian cattle via deep metric learning." Computers and Electronics in Agriculture 185 (2021): 106133.

[33] Ajitesh Kumar "Micro-average, Macro-average, Weighting: Precision, Recall, F1-Score", Analytics Yogi, Dec 2023, Accessed on 12 Jan 2025, Available: https://vitalflux.com/micro-average-macro-average-scoring-metrics-multi-class-classification-python/#google_vignette

*Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*